


nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



Competitive growth
maintains pecking order
in meerkat groups **PAGE 532**

COMPARE THE MEERKATS

REPRODUCIBILITY

THE INSIDE STORY

Survey reveals majority have
failed to replicate results

PAGE 452

HEALTH SCIENCE

CUTTING-EDGE MEDICINE

Why surgery is the next big
thing in type 2 diabetes


PAGE 459

CLIMATE

FROM A CLEAR SKY

Aerosol particles form
without pollution

PAGES 478, 521 & 527

 NATURE.COM/NATURE

26 May 2016 £10

Vol. 533, No. 7604



THIS WEEK

EDITORIALS

POSTDOCS More pay but fewer jobs on the way **p.438**

WORLD VIEW Treat antibiotic resistance as an ecological crisis **p.439**



DRONES Tiny flying robots with power to stick around **p.441**

Reality check on reproducibility

A survey of Nature readers revealed a high level of concern about the problem of irreproducible results. Researchers, funders and journals need to work together to make research more reliable.

Is there a reproducibility crisis in science? Yes, according to the readers of *Nature*. As we report on page 452, two-thirds of researchers who responded to a survey by this journal said that current levels of reproducibility are a major problem.

The ability to reproduce experiments is at the heart of science, yet failure to do so is a routine part of research. Some amount of irreproducibility is inevitable: profound insights can start as fragile signals, and sources of variability are infinite. But, the survey suggests, there is a bigger issue — and something that needs to be fixed. One-third of the survey respondents said that they think about the reproducibility of their own research daily, and more than two-thirds discuss it with colleagues at least monthly. The survey, of course, probably attracted researchers most interested in these issues. But it would be foolish to pretend that there is not serious concern.

What does ‘reproducibility’ mean? Those who study the science of science joke that the definition of reproducibility itself is not reproducible. Reproducibility can occur across different realms: empirical, computational and statistical. Replication can be analytical, direct, systematic or conceptual. Different people use reproducibility to mean repeatability, robustness, reliability and generalizability.

Economists and social scientists often use the term to mean that computer code and data are available so that someone would be able, if so inclined, to redo the same analysis using the same data. For bench scientists, who made up most of our respondents, it usually means that another scientist using the same methods gets similar results and can draw the same conclusions. We asked respondents to use this definition.

Even with a fixed definition, the criteria for reproducibility can vary dramatically between scientists. Senior scientists will not expect each tumour sample they examine under a microscope to look exactly like the images presented in a scientific publication; less experienced scientists might worry that such a result shows lack of reproducibility.

Scientists will need more rigorous use of terminology to get to grips with the problem. For now, broad-brush discussions and solutions are useful. Researchers lament that experiments that cannot be repeated do not give a solid foundation to build on.

Pressure to publish, selective reporting, poor use of statistics and finicky protocols can all contribute to wobbly work. Researchers can also be hampered from building on basically solid work by difficult techniques, poorly described methods and incompletely reported data. Funding agencies and publishers are helping to reduce these problems. Funders have changed their grant requirements and awarded grants for the design of courses to improve statistical literacy; journals are supporting technologies and policies that help to address inadequate documentation. For example, *Nature*’s Protocol Exchange website is available to host a protocol for any experiment, pre- or post-publication.

One-third of survey respondents report that they have taken the initiative to improve reproducibility. The simple presence of another

person ready to question whether a data point or a sample should really be excluded from analysis can help to cut down on cherry-picking, conscious or not. A couple of senior scientists have set up workflows that avoid having a single researcher in charge of preparing images or collecting results. Dozens of respondents reported steps to make better use of statistics, randomization or blinding. One described

“The criteria for reproducibility can vary dramatically between scientists.”

an institution-level initiative to teach scientists computer tools so they could share and analyse data collaboratively. Key to success was making sure that their data-management system also saved time. Another respondent spent three months working on a set of tools that enables different researchers to apply the same equations across different software and

computing environments and found that it led to praise, publications and collaborations.

Nature’s survey was launched before the US National Institutes of Health revised its grant requirements to improve reproducibility, and no survey questions asked explicitly about how research institutions might contribute, or how much time and money respondents would be willing to allocate to dedicated efforts to enhance reliability or replicate work. Our respondents seemed in principle receptive to such initiatives, which is encouraging for those — including *Nature* — who have already introduced steps to improve reproducibility. More steps are needed — starting with a discussion in the research community on how to properly credit, and talk to each other about, attempted replications. ■

Source material

Geneticists and historians need to work together on using DNA to explore the past.

Who brought down Rome? Few questions vex historians as much as the identity of the invaders who transformed the last vestiges of the great empire into a series of warring medieval territories. Was it long-distance migrants, the infamous barbarian hordes? Or was it diverse, local militias who moved to fill the power vacuums left by the diminished capital? Both?

This is not a question typically asked in these pages — historians have their own meetings and journals, after all. But as scholars continue to discuss the past, a new breed of scientists is trying to muscle in on the work of the present. These researchers want to use modern genetic techniques to answer historical questions, and as they do so, they are firmly treading on the toes of their colleagues in the

humanities. These geneticists promise answers: using analysis of DNA to discover what ‘really’ happened during the Bronze Age and the Viking sagas and replace ‘biased’ histories with cold, hard data.

Not all historians are embracing this new world. Many such studies, they complain, take a ‘sequence first, historicize later’ approach, in which researchers discover some shift in the genetic make-up in the inhabitants of a region, for example, and then postulate a historical event that might be responsible for the demographic change.

Some historians and linguists felt uneasy about papers published in this journal last year that found similarities between the genomes of people living on the Russian steppe 5,000 years ago and in Western Europe 4,500 years ago. The studies speculated that this correlation was the result of a massive migration to Europe of steppe people who also imported Indo-European languages, a family that includes nearly every dialect spoken on the continent (see *Nature* **522**, 140–141; 2015).

So, one might expect historians to be hostile to the latest sequencing effort. It aims to analyse DNA from 1,100 sets of ancient remains from across Italy, Austria, Hungary and the Czech Republic, to work out who filled the void left by the fall of the Roman Empire — or at least how the empire turned into the Lombard kingdom, which ruled parts of Italy between the sixth and eighth centuries AD.

Yet among the project leaders is a card-carrying medieval historian. Patrick Geary at the Institute for Advanced Study in Princeton, New Jersey, has shaped the questions that the project will tackle and how they will be asked. His colleagues must fight for the soul of their field before it is cannibalized, Geary argues. “If historians do not get involved and engage with this technology seriously, we’re going to see more and more studies that are done by geneticists with very little input from historians, or from frankly second-rate historians,” he says.

This week, he will lead a workshop that will gather 20 or so early-career historians and archaeologists at the Max Planck Institute for the

Science of Human History in Jena, Germany, to learn about ancient DNA and other quantitative tools that are disrupting how scholars probe the past.

Among the issues niggling at historians is the concern that an individual’s genetic make-up might be used interchangeably with his or her ethnic identity. Historians prefer to see ethnic groups, such as Anglo-Saxons or Franks, as fluid categories that involve identifying with one group while rejecting others. As such, the Lombard sequencing effort will not use DNA to define a genetic profile of the kingdom’s founders, but to ask nuanced questions about migration, continuity between earlier and later inhabitants, and whether their ancestry relates to how and where they were buried.

Other efforts to get geneticists and historians speaking the same language are under way. A consortium led by ancient-DNA researcher Hannes Schroeder, at the University of Copenhagen, recently won a €1.2-million (US\$1.3-million) grant for a collaborative research project called CITIGEN to make his field more accessible to historians and other humanities scholars. Like Geary, Schroeder worries that historians will be left behind if they fail to incorporate genetics into their research. “The train is running, and you jump on it or you miss it,” says Schroeder, who is also involved with an effort using ancient DNA to study the transatlantic slave trade.

The young historians and archaeologists who will get their first taste of molecular genetics this week will hopefully come away with a new tool to bring to their research. But they should be prepared — not just to understand genetics enough to read a paper, but to challenge insights gleaned with ancient DNA and to shape how the technology is used to interpret the past. After all, there are barbarians at the gates. ■

“Historians will be left behind if they fail to incorporate genetics into their research.”

Crunch time

Overtime pay for postdoctoral scientists is welcome — but could mean fewer positions.

Low pay and dwindling prospects of a permanent position have left many postdoctoral scientists feeling unloved. Yet last week, postdocs received appreciation from an unusual place: the US Department of Labor. In a long-overdue revision of the country’s overtime regulations, the department explicitly included postdocs among those who are eligible for overtime pay if they earn less than US\$47,476 per year. As we report on page 450, rather than pay overtime, many funders and universities are expected to raise the minimum wage for postdocs above that threshold.

The regulations are not perfect. They leave out those whose main responsibility is teaching, and the 1 December 2016 deadline to comply is tough for labs that operate on long-term budgets keyed to multi-year grant cycles. And the overtime threshold, which may become the de facto minimum pay for postdocs, still fails to meet the \$50,000 per year minimum recommended in a 2014 report on the biomedical workforce by the US National Academies.

Many established scientists look back on their postdoc wistfully as a time of unparalleled focus on research. Yet the postdoc now too often gives way to the ‘permado’. Postdocs may languish in that position for more than a decade, sometimes bouncing from one position to another. Their careers are in stasis even as their lives march on. Today’s postdocs are older than ever. They raise families and care for elderly parents. Many can hardly be considered trainees: they are functioning as lab managers or staff scientists, but are paid at a lower rate.

The stagnation comes because the number of academic faculty positions has not kept pace with the swelling postdoc ranks — a reality that is now receiving more attention, thanks in part to the laudable efforts of a cadre of established scientists who have made it their mission to address the postdoc plight. Francis Collins, head of the US National Institutes of Health (NIH), joined their ranks last week, when he announced plans to raise the pay for some NIH-funded postdocs to match the new overtime threshold. Other funding agencies should do the same.

Such changes do not come without trade-offs. The NIH budget is finite and higher postdoc salaries, however funded, are likely to translate into fewer postdoc positions — a consequence that worries the US National Postdoctoral Association in Washington DC. It also concerns principal investigators already struggling under flat research budgets.

But the change is needed. Principal investigators should take a hard look at their own labs and hiring practices. Do they need so many postdocs? A bigger lab does not necessarily mean greater impact.

Even graduate students can help to ease the postdoc glut. Many do not think hard about their own careers until they are well into their studies. Postdoc positions are so abundant — because they are cheap — that they have become the default career choice even for graduate students who have begun to doubt that they want to continue in science.

Graduate students should be encouraged to prepare earlier for careers outside academia. For example, the University of Massachusetts Medical School in Worcester has gone beyond the standard ‘alternative’ career seminars and made career preparation a mandatory part of the curriculum, with required workshops held periodically throughout a graduate student’s education. Students initially grumbled at being asked to spend more time away from the laboratory. By the end of the programme, 92% of them said they are glad that they did.

Such changes can go far to bring about reform — not just in the United States, but around the postdoc world. ■



Society must seize control of the antibiotics crisis

Pressure from the public could force firms to develop new drugs that treat resistant infections, says Carlos Amábile-Cuevas.

What are we to do about antibiotic resistance? Last week, another government report repeated stark warnings about the crisis, and offered some suggestions to improve the situation. The UK report, prepared by a panel chaired by the economist Jim O'Neill, naturally focused on financial incentives, including US\$1-billion prizes for pharmaceutical firms that develop new antimicrobial drugs (see go.nature.com/a8auos).

O'Neill, who in a previous job coined the term BRIC for the fast-growing economies of Brazil, Russia, India and China, suggested a different approach. As well as rewards for companies that invent new antibiotics, his report suggests punishments for those that do not try. Such firms, he writes, should pay a small fraction of annual sales into a fund to support rivals that invest in antibiotic research.

This is a welcome idea, but O'Neill does not go far enough. For too long, government moves to address the antibiotic-resistance crisis have focused on lucrative incentives: patent extensions, market exclusivity and higher prices. These mainly work to transfer public money into private hands, much in excess of what the research and development (R&D) actually costs. While we wait and see whether any of these interventions work, bacterial resistance continues to grow and spread, causing illness and death worldwide.

We need to take O'Neill's idea of a punitive levy and build on it. When it comes to the pharmaceutical industry and antibiotics, we need more sticks and fewer carrots.

Antibiotics are not like other drugs. The medical effects of prescribing and taking them are not restricted to one patient. In the words of the scientist Stuart Levy — one of the first to raise the public alarm over bacterial resistance — antibiotics are “societal drugs”. This societal impact justifies an approach to the development, marketing and use of antibiotics that is different from those of other medicines and consumer goods.

Ideally, governments would wield the sticks that would encourage this different approach — for example, by delaying or denying the approval of ‘me-too’ drugs from companies that do not invest in antibiotic research. That seems unlikely, but society can step in and act instead.

Take the agricultural overuse of antibiotics. Despite repeated warnings about the impact on human health, governments have failed to act. But consumer and campaign groups have had some success in pressuring fast-food chains not to buy meat from antibiotic-fed animals. Last year, McDonald's pledged to phase out meat from chickens treated with certain antibiotic drugs in its US restaurants by 2018.

This tactic of public pressure can be scaled up. Inaction that privileges other business over public interests has already been overturned in this way — people who choose wood and fish products labelled by the Forest Stewardship Council and Marine Stewardship Council are

rewarding companies that commit to public interests, and punishing those that do not.

We need to see antibiotics as another natural resource that demands careful stewardship. Most of these drugs are natural microbial products. Fundamentally, our ability to kill bacteria with old or new drugs depends on the microbes' natural susceptibility, which is non-renewable. Just like other ecological problems, antibiotic resistance is fuelled in part by the reckless behaviour of companies and a feeble response by regulators.

If the problem is the same, then perhaps we can borrow some ideas from the environmental movement to tackle antibiotic resistance.

Let's consider a ‘No Antibiotics, No Business’ initiative — NANBU for short. A campaign could issue a positive or negative NANBU rating to a pharmaceutical company depending, for example, on whether it invests

in R&D of genuinely new antibiotics. Other factors could be whether it promotes unethical or unwarranted clinical use of antibiotics (for sinusitis and bronchitis, for example) and if it sells antibiotics for agricultural use beyond the treatment of sick animals.

Physicians, then, could choose to prescribe drugs from NANBU-certified firms. Consumers could do the same when they purchase medicines over the counter, and press their doctors to prescribe treatments from certified companies. As sales across their product range dropped, firms that did not invest in antibiotic research would feel a compelling financial — and shareholder — pressure to do so.

If successful, such an initiative could address a societal problem with societal action. Giving people a way to become involved — by making informed choices — would also boost their sense of individual responsibility. This in turn could address the various forms of household antibiotic abuse, including a failure to complete a course of treatment.

To work, such a scheme would have to be kept out of the hands of governments and the companies themselves. It would need to be transparently managed, and more health-care professionals and consumers would need to be informed about what is at stake.

Just as with other societal problems, the extreme language of policy-makers on the severity of the antibiotic crisis is not matched by action. It is time to seize control for ourselves. Infections really do affect everybody. They are not ‘lifestyle’ conditions like some diseases. They affect young and old, rich and poor — and more people will die as more multiresistant bacteria emerge. Drug companies and governments have not been up to the task. It is time for society to act. ■

Carlos Amábile-Cuevas is director of the Lusara Foundation in Mexico City, a non-profit research institute that focuses on antibiotic resistance. e-mail: carlos.amabile@lusara.org

WHEN IT
COMES TO THE
PHARMACEUTICAL
INDUSTRY AND
ANTIBIOTICS,
WE NEED
MORE STICKS
AND FEWER CARROTS.

RESEARCH HIGHLIGHTS

Selections from the
scientific literature

PHYSICS

Precise clocks synced by lasers

Researchers have synchronized two optical clocks to a record-breaking level of accuracy.

Jean-Daniel Deschênes and his colleagues at the US National Institute of Standards and Technology in Boulder, Colorado, created optical clocks that keep accurate time with pulses of light. They used lasers to synchronize two such devices placed on a rooftop in the open air, and used mirrors to create a 4-kilometre path for the light. Despite atmospheric turbulence and even light snow, the clocks never deviated from each other by more than 40 femtoseconds over 50 hours. On average, they remained within 1 femtosecond of each other for more than 108 minutes.

The scheme could allow optical clocks to be used in gravitational and relativity experiments, for measuring quantum systems, and in ultra-precise global positioning networks, say the authors.

Phys. Rev. X 6, 021016 (2016)

CLIMATE CHANGE

Warming will hit the poorest first

As the climate warms over the coming decades, the poorest 20% of the world's population will see frequent temperature extremes sooner than the richest 20%.

Luke Harrington at Victoria University of Wellington and his colleagues used climate models to simulate the effect of rising levels of atmospheric carbon dioxide on daily temperature extremes for the rest of this century. Low latitudes, where most of the world's poorest people live,

will experience these changes in climate first. This is largely because these regions have less natural variability in temperature than mid-latitude regions, which are home to more of the world's wealthy.

Moving to a low-carbon economy will help poor communities the most, the authors say.

Environ. Res. Lett. 11, 055007 (2016)

ECOLOGY

Native insects embrace invader

An invasive plant has been gradually folded into an ecosystem's food webs.

Menno Schilthuizen at the Naturalis Biodiversity Center

in Leiden, the Netherlands, and his colleagues sampled insects from native bird cherry trees (*Prunus padas*) and exotic black cherry trees (*Prunus serotina*) in a Dutch national park. They found that the non-natives had around one-quarter of the number of insects on them, but almost twice the species diversity, compared with the native trees. The team also looked at preserved leaf specimens and found that the proportion of insect-eaten bird cherry leaves has remained stable at about 35% over the past 170 years, but that the proportion of invasive black cherry leaves consumed has increased from 18.8% to 40.6%.

This adaptation could slow the exotic plant's aggressive

spread — and efforts to control this by removing a proportion of the population may delay this process, the authors say.

PeerJ 4, e1954 (2016)

MICROBIOLOGY

Enzymes bust bacterial biofilms

Enzymes that break down tough films of disease-causing bacteria could one day be used as drugs.

Biofilms protect bacteria from antibiotics and are difficult to eradicate. To look for biofilm-fighting molecules, Lynne Howell at the Hospital for Sick Children in Toronto, Canada, and her colleagues studied



PHOTO SHOT/ALAMY

ANIMAL BEHAVIOUR

Gift helps spider to escape cannibalism

Male spiders use courtship gifts as shields to avoid being eaten by aggressive females.

Prior to mating, male nursery-web spiders (*Pisaura mirabilis*; male pictured on right, female on left) catch a prey item, wrap it in a silken package and present the 'nuptial gift' to the female. To find out why, Søren Toft and Maria

Albo of Aarhus University in Denmark presented female spiders with males in the lab. They found that males bearing gifts were never cannibalized before mating, whereas 19% of females ate males without gifts. How hungry a female was did not significantly influence male survival rates.

Biol. Lett. 12, 20151082 (2016)

Pseudomonas aeruginosa, a pathogen that mainly infects hospital patients. Several sugars produced by *P. aeruginosa* form components of its biofilm matrix, and the team identified two glycoside hydrolase enzymes that target and break down two of these sugars. When added to biofilms of clinical and environmental strains of the bacterium in culture, the enzymes degraded the films by 58% to 94%. They also inhibited biofilm formation for up to 72 hours and, when combined with an antibiotic, reduced bacterial growth by more than the antibiotic alone.

The authors say they have begun tests in animals to study the enzymes' therapeutic potential.

Sci. Adv. 2, e1501632 (2016)

EVOLUTION

Songs drove sunbird evolution

In two bird populations, differences in social traits, rather than just physical ones, are enough to generate new species.

Jay McEntee at the University of California, Berkeley, and his colleagues set out to understand how two species of sunbird that live side by side in East Africa, *Nectarinia moreaui* (pictured) and *Nectarinia fuelleborni*, evolved into separate species. The team compared the animals' genetics, physical traits, habitat preferences and songs, which are used by males to compete for territory and possibly for mates. They found that the species are physically similar and prefer the same types of food and habitat. However, the birds' songs differ in

duration and structure.

Genetic predispositions for different songs may have split the two species even though the birds share common habitats, the authors say.

Evolution <http://doi.org/bhpx> (2016)

PALAEOLOGY

Ancient origins of multicellular life

Large, multicellular life forms may have appeared on Earth one billion years earlier than was previously thought.

Macroscopic multicellular life had been dated to around 600 million years ago, but new fossils suggest that centimetres-long multicellular organisms existed as early as 1.56 billion years ago. Maoyan Zhu at the Chinese Academy of Sciences in Nanjing and his colleagues report the discovery of well-preserved fossils from northern China showing organisms up to 30 centimetres in length. The creatures' cells measure 6–18 micrometres in diameter and are closely packed. From comparisons with modern organisms, the authors suggest that the fossils were probably photosynthetic eukaryotes similar to modern algae.

The finding challenges the idea that this period of evolution was relatively uneventful, the authors say.

Nature Commun. 7, 11500 (2016)

REGENERATION

Muscle stem cells show dual purpose

Researchers have observed stem cells in the muscles of live zebrafish dividing to produce both more stem cells and cells that repair injury.

This 'asymmetric division' of stem cells has been observed in culture. To find evidence in living organisms, Peter Currie and his colleagues at Monash University in Clayton, Australia, engineered zebrafish larvae

SOCIAL SELECTION

Popular topics on social media

Scientific sceptics hit back

Champions of rational, evidence-based thinking are seething after a public rebuke at one of their own conferences.

Speaking on 15 May at the Northeast Conference on Science and Skepticism in New York, science journalist John Horgan said that sceptics — researchers and other people who promote the scientific method — spend too much time debunking 'soft' targets such as homeopathy when they should be going after tougher, 'hard' issues, such as whether regular mammograms save lives. Whereas some attendees welcomed the message, conference co-organizer Steven Novella, a neurologist at Yale University School of Medicine in New Haven, Connecticut, argued on his NeuroLogica blog that sceptics have been grappling with both hard and soft targets for years: "We are already miles past the superficial

➔ **NATURE.COM**
For more on popular papers:
go.nature.com/eaamst

framing that Horgan gives" Horgan said that he had an "impressionistic view" of the topics that were important to sceptics but that he hadn't taken a full survey.

to produce fluorescent proteins in their muscle cells, and monitored them over time using microscopes. They found that after an injury, stem cells infiltrated the wound and divided to produce two distinct cell types: one went on to form new muscle tissue, while the other had high levels of *cmet*, a marker specific to stem cells in muscle and other tissues.

This system of cell renewal could be a feature of vertebrates in general, the team says.

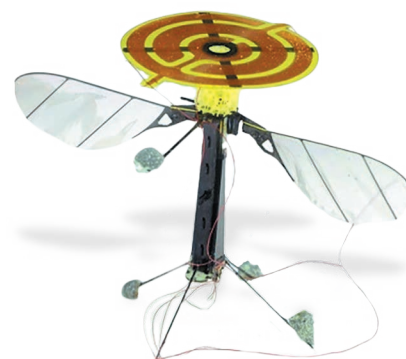
Science <http://dx.doi.org/10.1126/science.aad9969> (2016)

ROBOTICS

Robot hangs with electrostatic force

A lightweight flying robot can attach to and take off from objects in the environment by controlling electrostatic adhesion.

Moritz Graule and Robert Wood at Harvard University in Cambridge, Massachusetts, and their colleagues designed the insect-sized robot (pictured) to suspend from overhangs, such as those on trees and



buildings. The top of the device has a patch containing copper electrodes. Switching on the voltage between the electrodes induces opposing charges on the surface of a nearby target, generating an electrostatic attraction. Switching off the voltage releases the robot. The device can cling to a variety of surfaces, including glass, wood and leaves.

The authors suggest that this mechanism could help small aerial robots to conserve their energy and stay aloft for longer periods.

Science 352, 978–982 (2016)

➔ **NATURE.COM**
For the latest research published by Nature visit:
www.nature.com/latestresearch



AAAA

LOUIS A. HANSEN

SEVEN DAYS

The news in brief

BUSINESS

Myriad problems

Diagnostics firm Myriad Genetics, in Salt Lake City, Utah, is facing a legal challenge from people who say that the company refused to give them access to their own genomic data, thereby violating a US government rule on medical records. Myriad has now agreed to release the data, but the individuals filed the complaint to the US government on 19 May, in part to set the precedent that companies must legally provide the full results of genetic tests — not release them on a voluntary basis. The skirmish is the latest in a long-running war between Myriad and data-sharing advocates. See page 449 for more.

Bid for Monsanto

Chemical company Bayer has offered US\$62 billion to acquire the controversial agriculture giant Monsanto, of St Louis, Missouri. Bayer, headquartered in Leverkusen, Germany, announced the bid on 19 May and revealed the offer amount on 23 May in response to investor concern. If successful, the acquisition would create the largest agrochemical company in the

NUMBER CRUNCH

37%

The proportion of North America's 1,154 native bird species put on a 'watch list' in the first comprehensive assessment by the North American Bird Conservation Initiative on 18 May. The 432 species are most at risk of extinction and need urgent conservation action.

Source: www.stateofthebirds.org



JEAN SEBASTIEN EVYARD/AFP/GETTY

Fine over goat-cruelty allegations

The antibody producer Santa Cruz Biotechnology has been fined a record US\$3.5 million by the US Department of Agriculture (USDA) for alleged animal welfare violations. The fine is part of a 19 May settlement, after three USDA inspections said that the firm had kept goats and rabbits used for antibody production in cruel conditions.

The company, headquartered in Dallas, Texas, will also lose its licence to produce animals and research products. A January USDA inspection found no animals on the premises, suggesting that the firm had already stopped producing animals; thousands of goats and rabbits had disappeared. See go.nature.com/4wpcfm for more.

world. The development is the latest of several such attempted moves in the industry. In December 2015, US chemical giants Dow Chemical and DuPont announced that they would merge, and in February, the state-owned China National Chemical Corporation offered \$43 billion for the Swiss chemical and seed company Syngenta.

FUNDING

Postdoc pay

Minimum postdoctoral salaries in the United States are expected to rise as a result of a law change, finalized on 18 May by the US Department of Labor. The rule will make overtime pay mandatory for

many postdocs who are paid less than US\$47,476 per year. Overtime, which is paid at 1.5 times the normal hourly wage, kicks in once workers exceed 40 hours on the job in one week. But instead, funders and universities are likely to raise wage packets to meet that threshold. The average salary for a US postdoc is around \$45,000, with many earning substantially less. See pages 438 and 450 for more.

Pandemic funds

The World Bank Group has unveiled a new aid mechanism for developing countries that are dealing with infectious-disease outbreaks. The Pandemic Emergency Financing Facility, launched

on 21 May, and expected to be operational later this year, will provide up to US\$500 million to poor countries in the event of such outbreaks. The scheme was prompted by the delayed response to the Ebola outbreak in 2014, when the World Health Organization was unable to rally timely aid to fight the epidemic in West Africa. The aid packages will be funded through bond sales, insurance mechanisms and cash.

PEOPLE

Tech-prize first

Frances Arnold, a biochemical engineer at the California Institute of Technology in Pasadena, scooped the

Millennium Technology Prize for her work on directed evolution on 24 May. She is the first woman to win the €1-million (US\$1.1-million) prize, which is awarded every 2 years by the Technology Academy Finland. Arnold pioneered a technique for generating random DNA mutations to produce new proteins, which can then be tested and used in fields ranging from pharmaceuticals to renewable energy.

EVENTS

Australian quake

A magnitude-6.1 earthquake shook the Australian outback on 21 May, one of the largest in the country since seismic measurements started there in the early 1900s. It was a relatively rare instance of seismic activity in the middle of an otherwise stable plate of Earth's crust. It was also the biggest quake in Australia since a magnitude-6.2 tremor occurred off the western coast in 1997. There was no damage from the latest quake, which struck a sparsely populated region of the Northern Territory about 125 kilometres west of Uluru (Ayers Rock).

India space shuttle

India successfully launched a test version of its first reusable space shuttle (pictured) from Sriharikota, southern India,



on 23 May. The Reusable Launch Vehicle–Technology Demonstrator (RLV–TD) ascended to 65 kilometres before descending and re-entering the atmosphere at five times the speed of sound. It glided down to a defined landing spot over the Bay of Bengal, some 450 kilometres from Sriharikota. The reusable technology aims to reduce the cost of launching spacecraft to US\$2,000 per kilogram, about one-tenth of the current cost.

Environmental toll

Nearly one-quarter of deaths globally are the result of environmental effects, according to a report by the United Nations Environmental Programme on 23 May. It found that 12.6 million deaths in 2012 were attributable to a deteriorating environment, equivalent to 23% of total mortality. The biggest killer was air pollution, with poor-quality air contributing to about 7 million deaths each

year. Other environmental health risks included a lack of access to clean water and sanitation, exposure to chemicals, and natural disasters related to weather. Climate change was noted as an amplifier of negative health effects owing to its impact on land, water, biodiversity and weather.

French drug trials

France will tighten approval procedures for 'first-in-human' clinical trials, its health minister, Marisol Touraine, announced on 23 May. The announcement followed the release of a final report by the general inspectorate for social affairs on the death of a drug-trial participant in Rennes in January. The report reiterated its interim findings from February that although the trial protocol respected current regulations, there were major shortcomings surrounding the handling of

COMING UP

31 MAY – 2 JUNE

Representatives from science, business and government will gather in Berlin for the Group on Earth Observation's workshop for European projects.

go.nature.com/jrjwsx

4–7 JUNE

The biennial World Congress of Cardiology and Cardiovascular Health takes place in Mexico City. Research will cover everything from the effects of migration to obesity and health coverage.

go.nature.com/cev6tz

the fatal event by Biotrial, the trial contractor. Biotrial said that it "strongly contests" the findings, and claimed it had not been given a fair hearing.

RESEARCH

NFL rapped

Officials in the US National Football League (NFL) tried to influence how a donation to support sport-related health research was spent, according to a Congressional report released on 23 May. In 2012, the NFL gave an unconditional gift of US\$30 million to the US National Institutes of Health (NIH) to fund research, with a special focus on traumatic brain injuries. But an investigation by led by congressman Frank Pallone (Democrat, New Jersey) of the House of Representatives Energy and Commerce Committee says that the NFL pressured the NIH not to support a brain-trauma centre at Boston University in Massachusetts; after the NIH had awarded a \$16-million grant for the centre, the NFL reneged on this funding.

➔ NATURE.COM

For daily news updates see:

www.nature.com/news

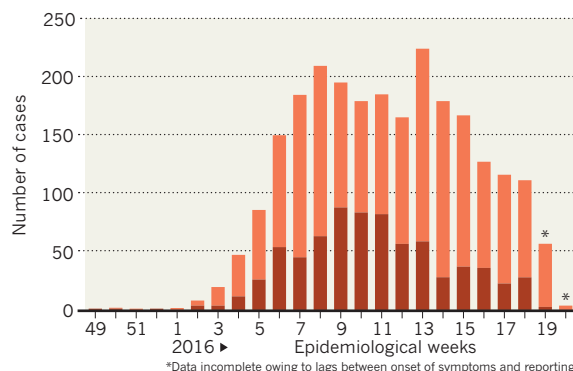
TREND WATCH

The World Health Organization announced on 19 May that the ongoing outbreaks of yellow fever in Africa constitute a serious public-health threat but not yet a Public Health Emergency of International Concern. The outbreak began last December in Luanda, Angola, but a delay in detection has enabled it to spread. As of 19 May, Angola had reported 2,420 suspected cases (736 lab confirmed) and 298 deaths. Travellers have taken the virus to the Democratic Republic of the Congo, China and Kenya.

YELLOW-FEVER OUTBREAK

The number of people contracting yellow fever in urban Angola seems to be on the wane.

■ Suspected ■ Confirmed



*Data incomplete owing to lags between onset of symptoms and reporting.

NEWS IN FOCUS

CLIMATE Billions of dollars needed for flotilla of carbon-sensing satellites **p.446**

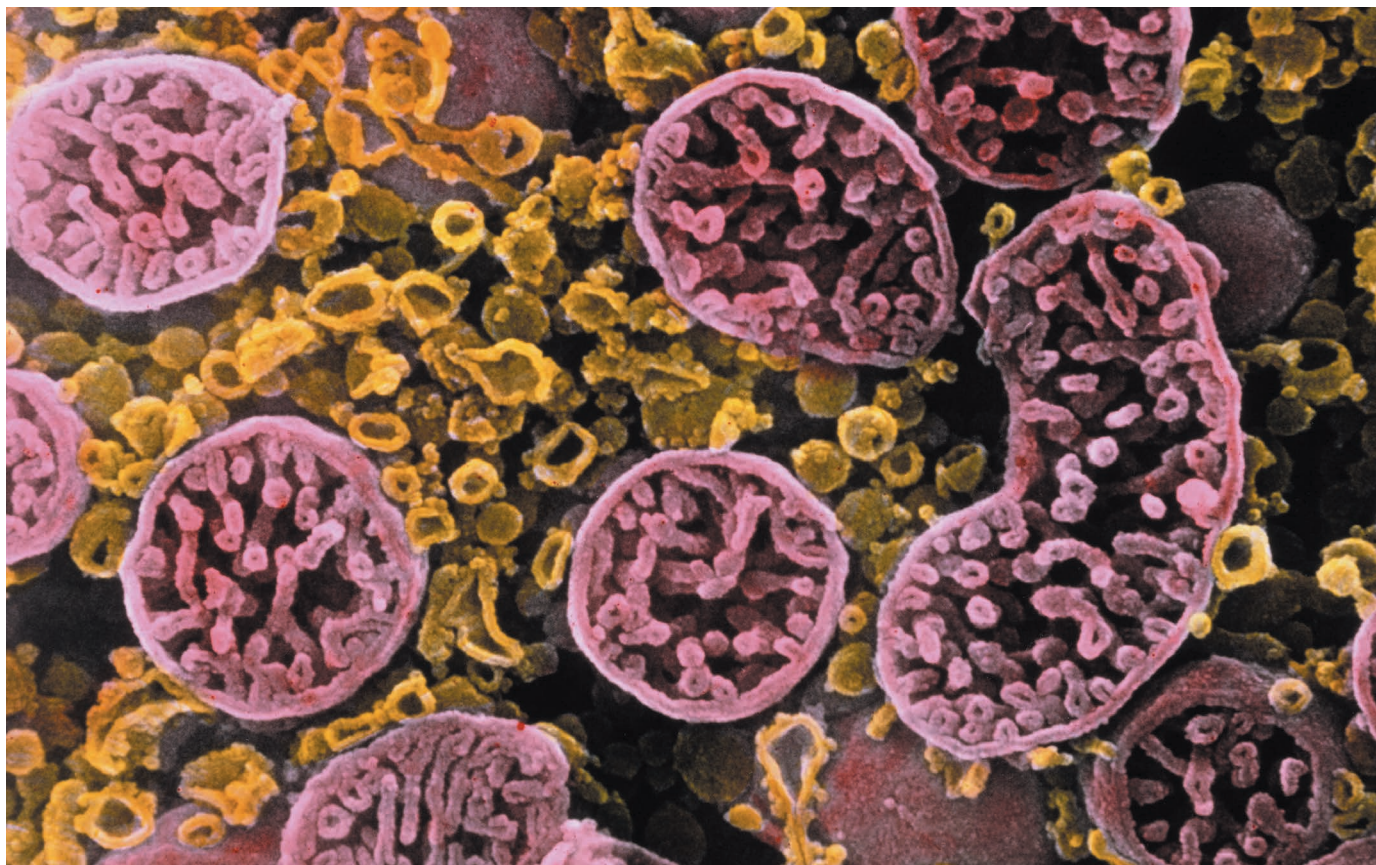
PHYSICS Silicon designs from Australia enter quantum-computer race **p.448**

INFECTIOUS DISEASE US reviews plan to infect mosquitoes with bacteria **p.450**



ARCHAEOLOGY Ancient toilets reveal what the Romans really did **p.456**

P. M. MOTTA/S. MAKABE/T. NAGURO/SPL



Mitochondria (shown in pink) are the cell's energy-producing structures and can contain harmful genetic mutations.

REPRODUCTIVE MEDICINE

Three-person embryos may not expel harmful genes

Technique to stop children inheriting mutated mitochondria has potential to backfire.

BY EWEN CALLAWAY

A gene-therapy technique that aims to prevent mothers from passing on harmful genes to children through their mitochondria — the cell's energy-producing structures — might not always work.

Mitochondrial replacement therapy involves swapping faulty mitochondria for those of a

healthy donor. But if even a small number of mutant mitochondria remain after the transfer — a common occurrence — they can outcompete healthy mitochondria in a child's cells and potentially cause the disease that the therapy was designed to avoid, experiments suggest.

"It would defeat the purpose of doing mitochondrial replacement," says Dieter Egli, a stem-cell scientist at the New York Stem Cell

Foundation Research Institute who led the work. Egli says that the finding could guide ways to surmount this hurdle, but he recommends that the procedure not be used in the meantime.

The UK government last year legalized mitochondrial replacement therapy, although the country's fertility regulator has yet to give the green light to its use in the clinic. In the United States, a panel convened by the National



► Academies of Sciences, Engineering, and Medicine has this year recommended that clinical trials of the technique be approved if preclinical data suggest that it is safe.

As many as 1 in 5,000 children are born with diseases caused by harmful genetic mutations in the DNA of their mitochondria; the diseases typically affect the heart, muscles and other power-hungry organs. Children inherit all their mitochondria from their mothers.

To prevent a mother who has harmful mitochondrial mutations from passing them to her children, the proposed remedy is to transplant the nuclear DNA of her egg into another, donor egg that has healthy mitochondria (and has been emptied of its own nucleus). The resulting embryo would carry the mitochondrial genes of the donor woman, and the nuclear DNA of the father and mother. These are sometimes called three-person embryos.

Current techniques can't avoid dragging a small number of the mother's mitochondria into the donor egg, totalling less than 2% of the resulting embryo's total mitochondria. This isn't enough to cause health problems. But researchers have worried that the proportion of faulty, 'carried-over' mitochondria may rise as the embryo develops. The UK Human Fertilisation and Embryology Authority (HFEA) — which will oversee clinical applications of mitochondrial replacement — has called for research into this possibility.

Egli's study offers some clarity (M. Yamada *et al. Cell Stem Cell* <http://doi.org/bhsj>; 2016). His team used eggs from women with healthy mitochondria, but otherwise

followed a procedure similar to the real therapy: transplanting nuclear DNA from one set of egg cells into another woman's egg cells. The team converted these eggs into embryos using two copies of the maternal genome instead of sperm (to discount any role for paternal DNA),

"I don't think it would be a wise decision to go forward with this uncertainty."

and the resulting embryonic stem cells at first harboured similarly minuscule levels. But one stem-cell culture showed a dramatic change: as the cells grew and divided, levels of the carried-over mtDNA jumped from 1.3% to 53.2%, only to later plummet back down to 1%. When the team split this cell line into different dishes, sometimes the donor egg's mtDNA won out, but in others, the carried-over mtDNA dominated.

COMPETING DNA

Exactly how the carried-over mitochondria rose to dominance is unclear. Egli suspects that the resurgence happened because one mitochondrion copied its DNA faster than the others could, which he says is more likely to occur when large DNA-sequence differences exist between the two populations of mitochondria.

Iain Johnston, a biomathematician at the University of Birmingham, UK, says that this theory makes sense. He was part of a team that found that, in mice with mitochondria from lab

extracted stem cells from the embryos and grew the cells in dishes in the lab. The embryos, on average, had just 0.2% of carried-over mitochondrial DNA (mtDNA),

strains and distantly related wild populations, one mitochondrial lineage tended to dominate (J. P. Burgstaller *et al. Cell Rep.* 7, 2031–2041; 2014). If mitochondrial replacement does reach the clinic, Johnston says that donors should be chosen such that their mitochondria closely match those of the recipient mother.

But Mary Herbert, a reproductive biologist at the University of Newcastle, UK, who is part of a team pursuing mitochondrial replacement, says that mitochondria behave very differently in embryonic stem cells than in normal human development. Levels of mutant mitochondria can fluctuate wildly in stem cells. "They are peculiar cells, and they seem to be a law unto themselves," she says. She calls the biological relevance of the latest report "questionable", and thinks that data from embryos cultured for nearly two weeks in the laboratory will provide more useful information than Egli's stem-cell studies.

An HFEA spokesperson says that the agency is waiting for further experiments on the safety and efficacy of mitochondrial replacement (including data from Herbert's team) before approving what could be the world's first mitochondrial replacement in humans.

Egli hopes that the HFEA considers his team's data. He thinks that the problem can be surmounted, for instance, by improving techniques to reduce the level of carried-over mitochondria or matching donors so that their mitochondria are unlikely to compete. Until this is shown for sure, he advocates caution. "I don't think it would be a wise decision to go forward with this uncertainty." ■

EARTH SCIENCE

Carbon-sensing satellite system faces high hurdles

Space agencies plan an advanced fleet, but technical and political challenges abound.

BY JEFF TOLLEFSON

Today just two satellites monitor Earth's greenhouse-gas emissions from space. But if the world's leading space agencies have their way, a flotilla of such probes could be launched beginning in 2030. The ambitious effort would help climate scientists to improve their forecasts — and it could also help to verify whether countries are upholding their commitments to reduce greenhouse-gas emissions.

But researchers will need to clear a daunting array of political and technical hurdles if they are to get the system — estimated to

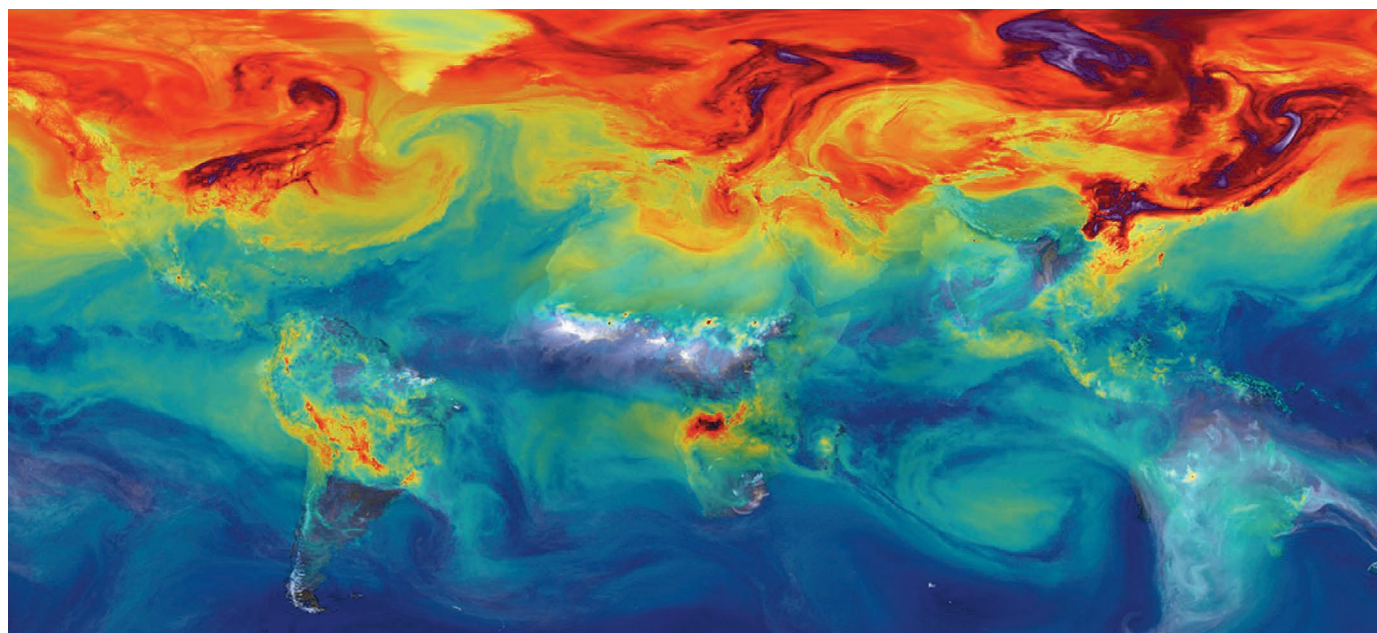
cost several billion dollars — off the ground. Competition for satellite launch slots is stiff: last year, for instance, the European Space Agency shelved plans for an advanced carbon-dioxide-monitoring probe in favour of a mission to measure plant growth. And scientists must still prove that satellite measurements of gases such as CO₂ and methane can match the accuracy of data from observatories on Earth.

"We have a small fleet of satellites that are being launched, but these are all just scientific experiments," says David Crisp, science team leader for NASA's Orbiting Carbon Observatory-2 (OCO-2). "What we are trying

to do now is just figure out how to monitor greenhouse gases from space."

Scientists have access to data from a pair of pioneering satellites: OCO-2, which launched in 2014 and measures CO₂, and Japan's Greenhouse Gases Observing Satellite (GOSAT), which launched in 2009 and tracks CO₂ and methane. NASA and the Japan Aerospace Exploration Agency are working to calibrate the instruments against each other and with a network of ground-based monitoring stations.

Both probes have a margin of error of about 0.5%, Crisp says. His team wants to reduce that to just 0.25% for the OCO-2 measurements.



A NASA atmospheric simulation from a computer model that traces how carbon dioxide emissions disperse in the atmosphere.

Meanwhile, nations are queuing up a new suite of satellites to lay the foundations for a larger monitoring effort (see ‘Counting carbon’). China will launch a pair of CO₂-monitoring satellites this year, and Japan plans to send up GOSAT-2 in 2018. NASA is looking ahead to OCO-3, which could launch as early as 2018. Unlike its predecessor, OCO-3 will not be a free-flying satellite, but a spectrometer built from OCO-2’s spare parts that will be installed on the International Space Station.

Crisp says that NASA scientists know how to build an instrument that is 20 times more powerful than OCO-2’s spectrometer, but that the agency does not have the funds to construct and launch it. “I can’t even compete for the money,” he says, “because it doesn’t exist.” At present, NASA is developing a satellite called ASCENDS that would use a laser to measure CO₂; unlike OCO-2, it would be able to track the gas at night and during winter at high latitudes.

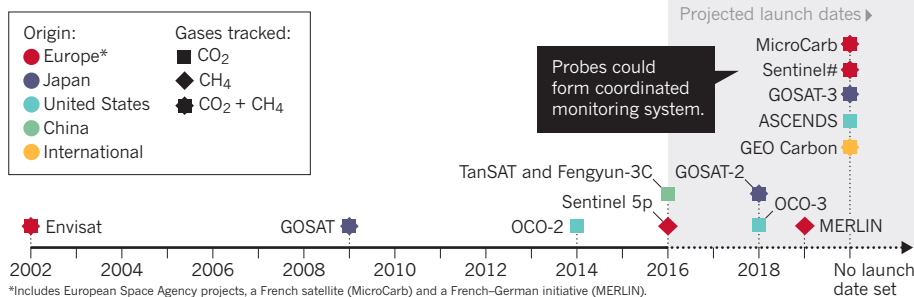
SHARPER VISION

Together, these satellites will extend a continuous record of CO₂ measurements that began in 2002 with SCIAMACHY, a spectrometer on the European Space Agency’s Envisat probe. The world’s space agencies want to transform this ad hoc system into a coordinated fleet of about six probes by 2030.

More than 60 agencies collaborated on a declaration finalized on 16 May that sets out a broad vision to develop and implement new monitoring technologies and to share the resulting data. France kick-started the effort last year, before December’s United Nations climate summit in Paris. At that meeting, nations adopted an agreement to curb heat-trapping emissions that also commits them to develop ways to verify whether they are

COUNTING CARBON

The number of satellites monitoring the world’s greenhouse-gas output could triple by 2030. Scientists are working to make the probes’ data on carbon dioxide and methane (CH₄) as accurate as those collected by observatories on Earth.



meeting their climate goals. The nascent satellite programme could help nations to fulfil that requirement.

“It’s very general, but it’s a beginning,” says Jean-Yves Le Gall, the head of the French space agency, CNES. He says that the group is discussing how to distribute and use data from the system before its meeting in Mexico in September.

The space agencies’ efforts are a welcome surprise to many researchers. “I think it’s very important,” says Heinrich Bovensmann, a remote-sensing specialist at the University of Bremen in Germany. “Now we have to see what really comes out of it.”

The current generation of satellites measure how much CO₂ is in a given column of air. But what scientists and governments really want to know is where that CO₂ came from. To pinpoint emissions produced by industrial activity, researchers must also determine how much CO₂ is absorbed and released by land and ocean ecosystems.

This requires combining atmospheric CO₂

measurements with accurate information about how the natural environment takes up and releases the gas, and plugging the data into detailed air-current models that can trace where the CO₂ probably came from. In essence, the task requires developing something akin to current weather forecast systems — but run in reverse.

Achieving the precision needed to verify whether nations are meeting their emissions goals is a tall order, says Stephen Pacala, a climate researcher at Princeton University in New Jersey who led a 2010 report on space-based carbon monitoring for the US National Academies of Sciences, Engineering, and Medicine. He says that satellites would need to be more accurate than current greenhouse-gas inventories, which are calculated using a variety of data on fossil-fuel consumption and economic and land-use trends.

But Crisp and his colleagues remain confident that they can deliver greenhouse-gas data that will be useful for both scientists and policymakers. “We’ve got training wheels on for the next decade,” he says. “Once we learn how to ride, we’ll build a bicycle.” ■

“We’ve got training wheels on for the next decade.”

PHYSICS

Silicon quantum computers take shape in Australia

Blueprints emerge from centre tasked with creating a practical quantum device.

BY ELIZABETH GIBNEY

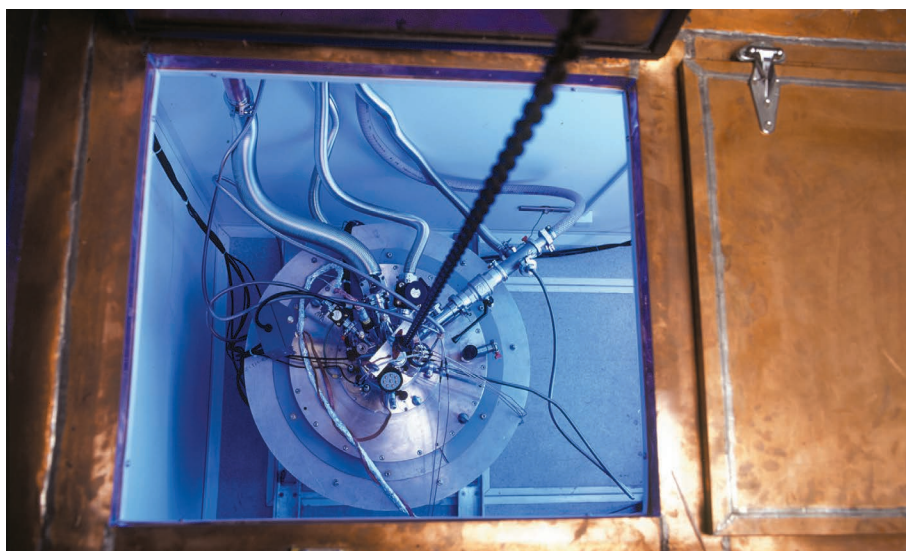
Silicon is at the heart of the multibillion-dollar computing industry. Now, efforts to harness the element to build a quantum processor are taking off, thanks to elegant designs from an Australian collaboration.

In July, the Centre for Quantum Computation and Communication Technology, which is based at the University of New South Wales (UNSW) in Sydney, will receive the first instalment of a Aus\$46-million (US\$33-million) investment. The money comes from government and industry sources whose goal is to create a practical quantum computer.

At an innovation forum in London on 6 May, hosted by *Nature* and start-up accelerator Entrepreneur First, two physicists from a group at the UNSW pitched a plan to reach that goal. Their audience was a panel of entrepreneurs and scientists, who critiqued ideas for commercializing a range of quantum technologies, including sensors, computer security and a quantum internet as well as quantum computers.

So far, the UNSW team has demonstrated a system with quantum bits, or qubits, only in a single atom. Useful computations will require linking qubits in multiple atoms. But the team's silicon qubits hold their quantum state nearly a million times longer than do systems made from superconducting circuits, a leading alternative, UNSW physicist Guilherme Tosi told participants at the event. This helps the silicon qubits to perform operations with one-sixth of the errors of superconducting circuits.

If the team can pull off this low error rate in a larger system, it would be "quite amazing", said Hartmut Neven, director of engineering at Google and a member of the panel. But he cautioned that in terms of performance, the system is far behind others. The team is aiming



The refrigerator at the University of New South Wales that is used to cool silicon quantum chips.

for ten qubits in five years, but both Google and IBM are already approaching this with superconducting systems. And in five years, Google plans to have ramped up to hundreds of qubits.

A second group from the UNSW has a less robust silicon design that has already demonstrated calculations that link up two qubits, a building block that paves the way for creating more-complex devices.

In a regular computer, each bit can be 'on' or 'off'. In a quantum computer, the qubits can be both on and off at once, which allows them to perform many computations in parallel. This should allow quantum computers to zip through calculations that would take a normal computer longer than the age of the Universe, although the best devices are still much too small to do so.

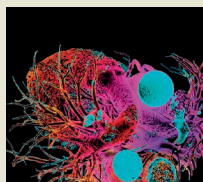
Silicon is an attractive base for a universal

quantum computer — one that can carry out any quantum algorithm — because it is potentially compatible with the microelectronics of existing computers. But silicon systems are still years behind their rivals. One issue has been how to keep delicate quantum states alive for long enough to perform operations.

The scheme showcased at the innovation forum by Tosi and fellow physicist Vivien Schmitt — who are both part of Andrea Morello's lab at the UNSW — addresses this challenge. The qubits are the spins of the electrons and nuclei in phosphorus atoms embedded in a silicon lattice, and are controlled using a special system of electric fields. Because the spins respond only to very specific, tuneable frequencies, they are robust to electrical noise. That allows the qubits to keep their quantum states for one minute and to operate perfectly


**MORE
ONLINE**

TOP STORY



Protective gene offers hope for next blockbuster heart drug
go.nature.com/vclfnl

MORE NEWS

- Why women earn less: Just two factors explain post-PhD pay gap
go.nature.com/jmflm7
- Winners and losers emerge in UK funding shake-up
go.nature.com/ntao9q
- Have physicists found a new force of nature?
go.nature.com/69bqyn

NATURE PODCAST



How trees help clouds form; a Neanderthal building project; and comparing meerkats
nature.com/nature/podcast

99.9% of the time, said Tosi.

Moreover, the electrically controlled qubits can communicate with each other at larger distances than can the qubits in other silicon designs. That bodes well for scaling up because the qubits can be far enough apart to allow room for control and read-out instruments to be placed between them. The atoms also do not need to be placed precisely, so they would fit with existing microprocessor-fabrication techniques, added Tosi.

Although Morello's team has demonstrated the high precision only in a single atom, the researchers have started to experiment with a two-atom system, and expect this level of performance to scale up. "From there, there are no basic first-principle barriers," said Tosi.

He has patented a design for a larger-scale computer and says that it should be possible to manufacture chips that use their system in plants similar to those already used to make present-day microprocessors.

SPIN CITY

Another silicon project from UNSW, led by physicist Andrew Dzurak, uses as its qubits the spins of electrons in a set-up that is based on modified electrical transistors. Although the qubits are less robust than those in the Morello design, Dzurak's team demonstrated two-qubit calculations last October.

At the London event, Tosi and Schmitt were grilled on the Morello team's plan. "It's a beautiful scheme," said John Morton, an electrical engineer at University College London. But he cautioned that aspects of the business model that the duo presented as part of the event seemed reliant on superconducting qubits not getting any better. "That's a risky thing to do," Morton said. Neven agreed, pointing out that silicon qubits will need to demonstrate advantages in other areas if they are to get a foothold in the future quantum-computer market.

Like superconducting qubits, silicon qubits must be kept at a fraction of a degree above absolute zero. Morton said that the Morello design's big advantage is that the qubits are atomic scale. In principle, that would allow many more to be placed on each chip than is possible with superconducting qubits, which are around 100 micrometres across. To seize this opportunity, the design needs to shed the bulky microwave devices that it proposes as its means to operate between distant qubits, said Neven.

The successes of the Australian quantum centre — which is led by UNSW physicist Michelle Simmons — are partly down to its wider expertise in nanoscale silicon fabrication, says Peter Knight, a physicist at Imperial College London and a panel member at the event. When it comes to fabrication, he says, the centre is way ahead of others working in silicon. ■

CANCER SCREENING

Myriad Genetics caught in data fight

Complaint to US government alleges that company violated people's right to access health information.

BY ERIKA CHECK HAYDEN

Genetic-testing firm Myriad Genetics is facing a legal challenge from people who say the company refused to give them access to their own genomic data, in violation of a US government rule on medical records.

Although Myriad has now agreed to release the data to those individuals, the patients are pressing ahead with their complaint to the US government. The skirmish is the latest in a long-running war between Myriad and data-sharing advocates, and it could ultimately force the company to provide genetic information that patients could then share with scientists.

The patients, who are represented by the American Civil Liberties Union (ACLU), filed the complaint on 19 May with the US government alleging that Myriad, of Salt Lake City, Utah, had declined to release complete results of tests for the genes *BRCA1* and *BRCA2*. Some variants of these genes are linked to higher risk of cancer; for others, the link to disease is unclear or the variants are considered to be harmless.

Myriad refused to report 'benign' *BRCA* variants back to patients when they requested this information in February. The ACLU says that the company's denial violates a rule released by the US government in January that gives patients the right to obtain their full lab test results under the Health Insurance Portability and Accountability Act.

TOTAL ACCESS

Breast-cancer survivor AnneMarie Ciccarella, one of the people who filed the complaint, said that she wants access to her complete data so that she can share it with scientists who are trying to understand the genetic contributions to cancer. "I want to see that the research community has access to every bit of data that has been generated from my body," she said.

On 18 May, after the ACLU announced a press conference to discuss the complaint, Myriad released the data that Ciccarella and her three co-complainants had requested. "We believe the complaint lacks merit and should not be accepted," the company said in a 19 May statement.

Ciccarella and the others who brought the

complaint are pressing ahead with their case, in part to set the precedent that companies must legally provide the full results of genetic tests — not release it on a voluntary basis.

Observers say that makes sense, especially given Myriad's history. The company had previously tried to block rivals from providing *BRCA* tests, asserting that it held patents that gave it the exclusive right to perform such diagnostics.

That changed in June 2013, when the US Supreme Court invalidated Myriad's patents after the ACLU mounted a legal challenge. Myriad has not shared its large database on thousands of *BRCA* variants, despite requests from researchers studying the genetics of breast cancer. But it may now be forced to provide

"I want to see that the research community has access to every bit of data that has been generated from my body."

individual results on a patient-by-patient basis if the government decides to accept the latest complaint.

"If I were the plaintiffs, I'd want to make sure the government said that Myriad had to do what it did," says lawyer and bioethicist Hank Greely at Stanford University in California. "If you're a consumer advocate in the health-care space, Myriad may not be a company you trust."

Heidi Rehm, a geneticist at the non-profit company Partners Healthcare Personalized Medicine in Cambridge, Massachusetts, drafted a statement of support for the complaint against Myriad. She says that as researchers learn more about genetic risks of cancer, they're finding that variants once considered benign might actually contribute to cancer risk.

Rehm and other researchers are pushing for companies and individuals to share their genetic test data with open databases such as ClinVar, and she says that the push for data sharing is gathering increasing momentum. US insurance company Aetna, for instance, has said that it will favour testing companies that deposit data in ClinVar. And the US Food and Drug Administration is considering whether to give companies incentives to deposit their tests results in the database. ■

Additional reporting by Heidi Ledford.

POLICY

US law may lift postdoc pay

New labour rules require overtime pay for many.

BY HEIDI LEDFORD

A change in US labour regulations will render many postdocs eligible for overtime pay — and create an incentive to raise their wages. The law may ultimately lead to fewer postdocs. But some say that the policy could spark much-needed changes to a research system that relies heavily on postdocs, yet offers them few opportunities for career advancement.

The new rule, finalized on 18 May by the US Department of Labor, will make overtime pay mandatory for many postdocs who make less than US\$47,476 per year. Overtime is paid at 1.5 times the normal hourly wage, and kicks in once workers exceed 40 hours on the job in 1 week.

But rather than pay the overtime, funders and universities are expected to raise minimum postdoc salaries to meet that threshold before the rule takes effect in December. “It’s a win for postdocs,” says Benjamin Corb, director of public affairs at the American Society for Biochemistry and Molecular Biology in Rockville, Maryland. “And I think it’s the right move for the community.”

The average salary for a US postdoc is around \$45,000, with many making substantially less. But as postdocs become more expensive, laboratories may begin to cut back on the number that they hire. “You can’t just say everybody’s going to get more money,” says Paula Stephan, who studies the economics of scientific research at Georgia State University in Atlanta.

But fewer postdocs, she says, may be what the system needs. In December 2014, the US National Academies published *The Postdoctoral Experience Revisited*, a report arguing that postdoc salaries should be raised to \$50,000 a year, and that many postdocs should be reclassified — and better paid — as staff scientists. In 2015, a poll of 20,000 *Nature* readers found that scientists are eager to see more permanent staff-scientist positions created. That change has been difficult to implement while postdoc salaries remain low.

Corb agrees that short-term cutbacks in postdoc numbers could yield a healthier research system: “To increase postdoc pay and thin out the pool of postdocs may end up, in the long run, being a net positive for the enterprise.” ■



Male *Aedes aegypti* mosquitoes infected with *Wolbachia* bacteria are unable to produce offspring.

INFECTIOUS DISEASE

Infected mosquitoes could fight Zika

US government reviews plan to use bacteria to reduce number of disease-carrying mosquitoes.

BY EMILY WALTZ

The United States could soon become the first country to approve the commercial use of a common bacterium to fight the spread of mosquitoes that can transmit viruses such as Zika, dengue and Chikungunya.

The US Environmental Protection Agency (EPA) is reviewing an application from the biotechnology start-up MosquitoMate to use the bacterium *Wolbachia pipientis* as a tool against the Asian tiger mosquito (*Aedes albopictus*). The company plans to market *Wolbachia* as a pesticide — one that kills only mosquitoes, and leaves other insects untouched. The EPA’s decision on the matter will come after a public-comment period that ends on 31 May.

MosquitoMate’s strategy involves rearing mosquitoes infected with a particular strain of *Wolbachia* and releasing the males into the environment. When these male mosquitoes mate with wild females who do not carry the

same strain of *Wolbachia*, the resulting fertilized eggs don’t hatch, because the paternal chromosomes do not form properly. As infected male mosquitoes continue to be released to breed with wild partners, the pest population dwindles.

Eight countries have now reported cases of microcephaly or other fetal birth defects that are probably caused by Zika, leading officials in many areas to consider new options for reducing mosquito populations. “We need as many effective tools as we can get, so we need to give *Wolbachia* a try,” says Tom Scott, an entomologist at the University of California, Davis. “That will require a well-developed plan for how trials would be done.”

MosquitoMate, which was started by researchers at the University of Kentucky in Lexington, has tested *Wolbachia* in *A. albopictus* mosquitoes in three states over the past three years. The approach has reduced the wild mosquito population by more than 70% in those areas, says Stephen Dobson, an entomologist at the University

of Kentucky and founder of the company.

MosquitoMate is also using *Wolbachia* to target the mosquito *Aedes aegypti*, which is thought to be the main vector for Zika. The firm began field trials this month of infected *A. aegypti* mosquitoes in Clovis, California, and has applied to conduct similar tests in Florida and in Orange County, California.

Other groups are also investigating *Wolbachia*'s ability to stamp out *A. albopictus*. Researchers from Sun Yat-sen University in Guangzhou, China, and Michigan State University in East Lansing began field trials of *Wolbachia*-infected mosquitoes last year on Shazai Island in Guangzhou.

In March, the tests expanded to Dadao Island, also in Guangzhou. The researchers are releasing 1.5 million male *A. albopictus* per week, with plans to increase that to 5 million per week by the end of August. "Our mosquito factory is currently the largest one in the world," says Zhiyong Xi, a medical entomologist and microbiologist at Michigan State, who oversees the project.

But such large, ongoing releases of mosquitoes may be too expensive for many countries or cities. With this in mind, the non-profit international collaboration Eliminate Dengue is testing an approach that requires the rearing of many fewer mosquitoes. Instead, it uses a starter set of mosquitoes that carry *Wolbachia* to infect an entire wild population.

The resulting offspring harbour the bacteria yet develop normally. But the *Wolbachia* infection prompts an immune response and consumes key cellular resources, which helps to prevent viruses — such as Zika and dengue — from growing and replicating in these mosquitoes (H. L. C. Dutra *et al. Cell Host Microbe* <http://doi.org/bhsh>; 2016).

"We're not trying to kill or suppress the mosquito population, we're just making them ineffective at transmitting a range of pathogens to people," says Scott O'Neill, head of Eliminate Dengue and dean of science at Monash University in Melbourne, Australia. The group is testing the technology against *A. aegypti* in open trials in Indonesia, Vietnam, Australia, Brazil and Colombia. O'Neill hopes to make the system affordable for developing countries at a cost of about US\$1 per person.

But using *Wolbachia* to suppress or infect mosquito populations has never been proved to reduce the incidence of Zika or dengue in humans. O'Neill says that the Eliminate Dengue group has seen "a collapse of dengue transmission" where it has released *Wolbachia*-infected mosquitoes. The group is trying to verify those observations with controlled, randomized studies that are now under way

"We need as many effective tools as we can get."

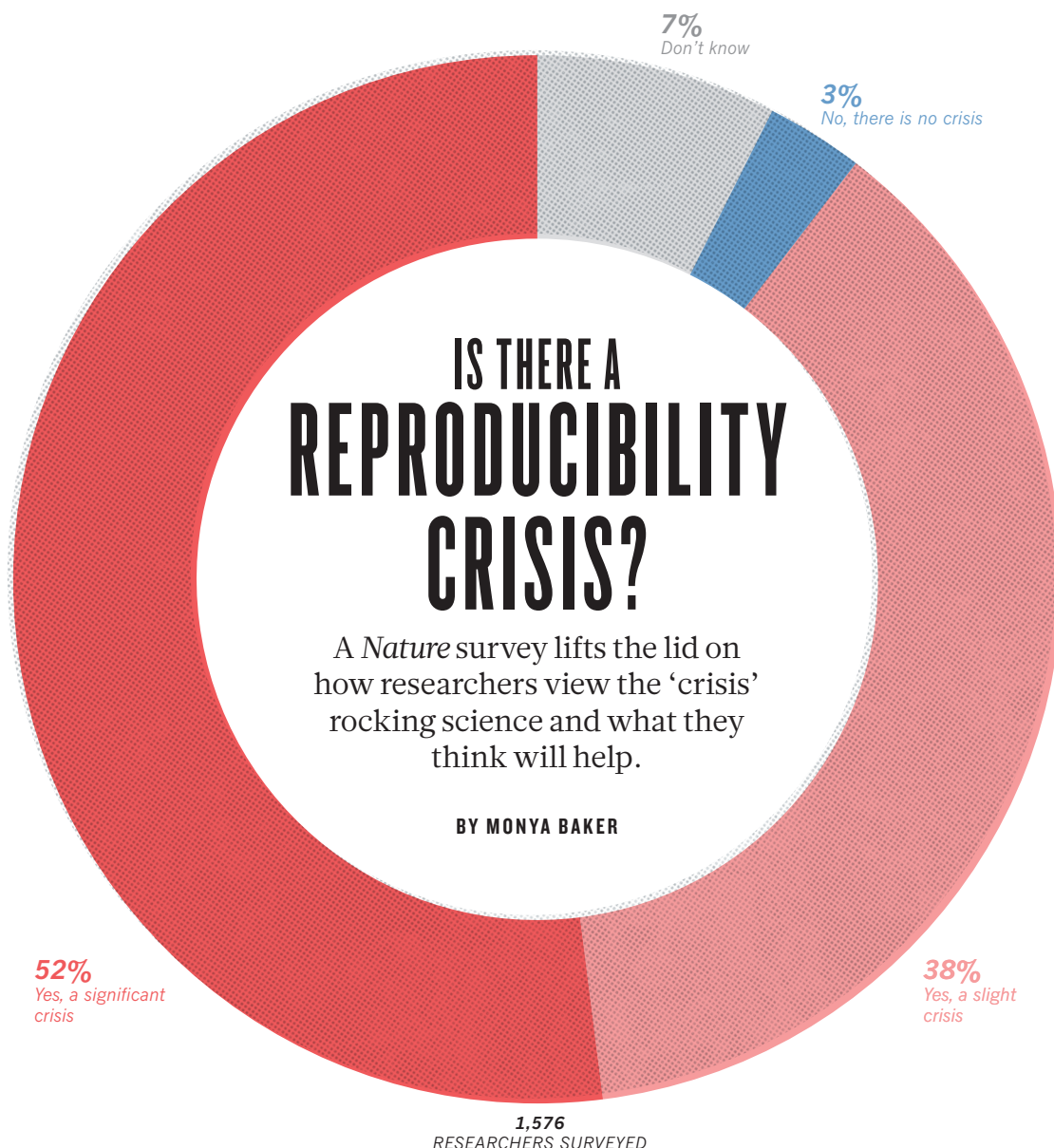
in Indonesia and Vietnam.

So far, tests of mosquitoes infected with *Wolbachia* have prompted little public resistance. By contrast, US residents have used yard signs, social-media campaigns and a petition to protest against proposed trials of genetically engineered mosquitoes developed by Oxitec of Milton Park, UK. Oxitec and MosquitoMate each alter male mosquitoes using a lethal reproductive weapon before releasing them into the environment to mate with and suppress their own kind. But Oxitec modifies its mosquitoes with a gene, whereas MosquitoMate uses a bacterium.

The US Food and Drug Administration, which is considering Oxitec's proposal for a field trial in Florida, received more than 2,600 public comments on the plan. But as *Nature* went to press, the EPA, which is accepting public input on MosquitoMate's plan until 31 May, had received just one comment. ■

CORRECTION

The News Feature 'The material code' (*Nature* **533**, 22–25; 2016) wrongly implied that the phrase 'materials genome' was invented solely by Gerbrand Ceder. It was independently invented and copyrighted by Zi-Kui Liu of Pennsylvania State University.



More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments. Those are some of the telling figures that emerged from *Nature's* survey of 1,576 researchers who took a brief online questionnaire on reproducibility in research.

The data reveal sometimes-contradictory attitudes towards reproducibility. Although 52% of those surveyed agree that there is a significant 'crisis' of reproducibility, less than 31% think that failure to reproduce published results means that the result is probably wrong, and most say that they still trust the published literature.

Data on how much of the scientific literature is reproducible are rare and generally bleak. The best-known analyses, from psychology¹ and cancer biology², found rates of around 40% and 10%, respectively. Our survey respondents were more optimistic: 73% said that they think that at least half of the papers in their field can be trusted, with physicists and chemists generally showing the most confidence.

The results capture a confusing snapshot of attitudes around these issues, says Arturo Casadevall, a microbiologist at the Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland. "At the current time there is no consensus on what reproducibility is or should be." But just recognizing that is a step forward, he says. "The next step may be identifying what is the problem and to get a consensus."

Failing to reproduce results is a rite of passage, says Marcus Munafo, a biological psychologist at the University of Bristol, UK, who has a long-standing interest in scientific reproducibility. When he was a student, he says, "I tried to replicate what looked simple from the literature, and wasn't able to. Then I had a crisis of confidence, and then I learned that my experience wasn't uncommon."

The challenge is not to eliminate problems with reproducibility in published work. Being at the cutting edge of science means that sometimes results will not be robust, says Munafo. "We want to be discovering new things but not generating too many false leads."

THE SCALE OF REPRODUCIBILITY

But sorting discoveries from false leads can be discomfiting. Although the vast majority of researchers in our survey had failed to reproduce an experiment, less than 20% of respondents said that they had ever been contacted by another researcher unable to reproduce their work (see 'A 'crisis' in numbers'). Our results are strikingly similar to another online survey of nearly 900 members of the American Society for Cell Biology (see go.nature.com/kbzs2b). That may be because such conversations are difficult. If experimenters reach out to the original researchers for help, they risk appearing incompetent or accusatory, or revealing too much about their own projects.

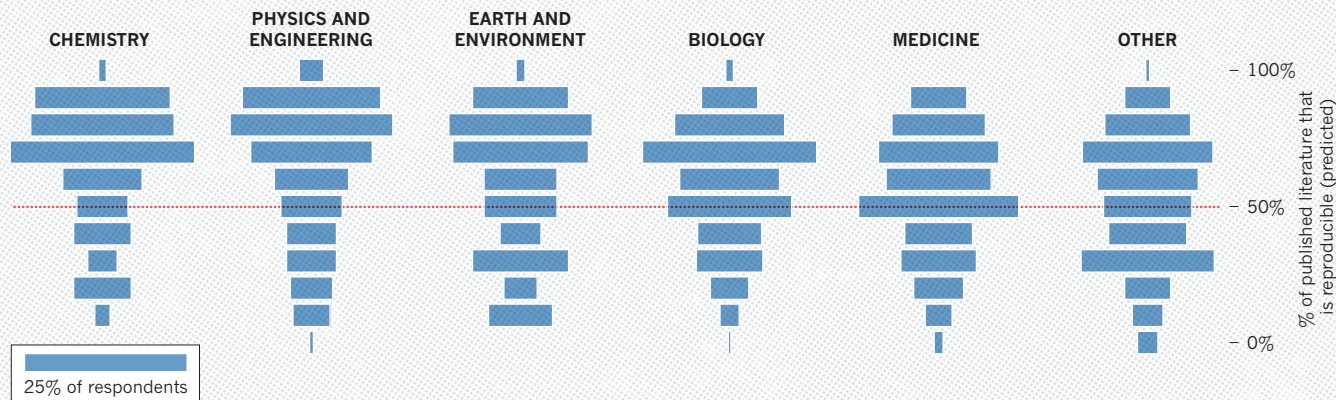
A minority of respondents reported ever having tried to publish

A 'CRISIS' IN NUMBERS

Nature surveyed 1,576 scientists online to get their thoughts on reproducibility in their field and in science in general. See go.nature.com/2vjr4y for more charts and access to the full data.

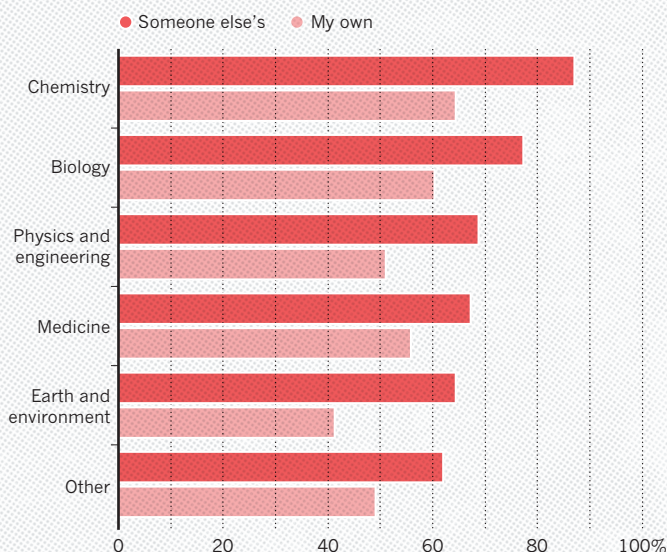
HOW MUCH PUBLISHED WORK IN YOUR FIELD IS REPRODUCIBLE?

Physicists and chemists were most confident in the literature.



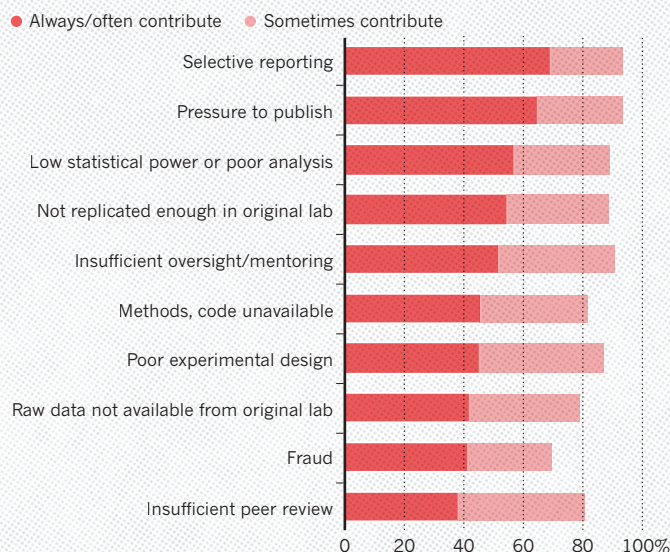
HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?

Most scientists have experienced failure to reproduce results.



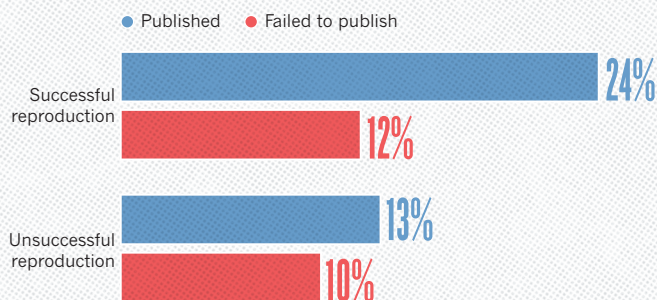
WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.



HAVE YOU EVER TRIED TO PUBLISH A REPRODUCTION ATTEMPT?

Although only a small proportion of respondents tried to publish replication attempts, many had their papers accepted.

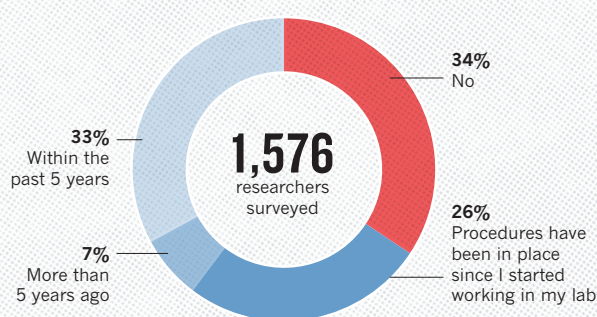


Number of respondents from each discipline:

Biology 703, Chemistry 106, Earth and environmental 95, Medicine 203, Physics and engineering 236, Other 233

HAVE YOU ESTABLISHED PROCEDURES FOR REPRODUCIBILITY?

Among the most popular strategies was having different lab members redo experiments.



a replication study. When work does not reproduce, researchers often assume there is a perfectly valid (and probably boring) reason. What's more, incentives to publish positive replications are low and journals can be reluctant to publish negative findings. In fact, several respondents who had published a failed replication said that editors and reviewers demanded that they play down comparisons with the original study.

Nevertheless, 24% said that they had been able to publish a successful replication and 13% had published a failed replication. Acceptance was more common than persistent rejection: only 12% reported being unable to publish successful attempts to reproduce others' work; 10% reported being unable to publish unsuccessful attempts.

Survey respondent Abraham Al-Ahmad at the Texas Tech University Health Sciences Center in Amarillo expected a "cold and dry rejection" when he submitted a manuscript explaining why a stem-cell technique had stopped working in his hands. He was pleasantly surprised when the paper was accepted³. The reason, he thinks, is because it offered a workaround for the problem.

Others place the ability to publish replication attempts down to a combination of luck, persistence and editors' inclinations. Survey respondent Michael Adams, a drug-development consultant, says that work showing severe flaws in an animal model of diabetes has been rejected six times, in part because it does not reveal a new drug target. By contrast, he says, work refuting the efficacy of a compound to treat Chagas disease was quickly accepted⁴.

THE CORRECTIVE MEASURES

One-third of respondents said that their labs had taken concrete steps to improve reproducibility within the past five years. Rates ranged from a high of 41% in medicine to a low of 24% in physics and engineering. Free-text responses suggested that redoing the work or asking someone else within a lab to repeat the work is the most common practice. Also common are efforts to beef up the documentation and standardization of experimental methods.

Any of these can be a major undertaking. A biochemistry graduate student in the United Kingdom, who asked not to be named, says that efforts to reproduce work for her lab's projects doubles the time and materials used — in addition to the time taken to troubleshoot when some things invariably don't work. Although replication does boost confidence in results, she says, the costs mean that she performs checks only for innovative projects or unexpected results.

Consolidating methods is a project unto itself, says Laura Shankman, a postdoc studying smooth muscle cells at the University of Virginia, Charlottesville. After several postdocs and graduate students left her lab within a short time, remaining members had trouble getting consistent results in their experiments. The lab decided to take some time off from new questions to repeat published work, and this revealed that lab protocols had gradually diverged. She thinks that the lab saved money overall by getting synchronized instead of troubleshooting failed experiments piecemeal, but that it was a long-term investment.

Irakli Loladze, a mathematical biologist at Bryan College of Health Sciences in Lincoln, Nebraska, estimates that efforts to ensure reproducibility can increase the time spent on a project by 30%, even for his theoretical work. He checks that all steps from raw data to the final figure can be retraced. But those tasks quickly become just part of the job. "Reproducibility is like brushing your teeth," he says. "It is good for you, but it takes time and effort. Once you learn it, it becomes a habit."

One of the best-publicized approaches to boosting reproducibility is pre-registration, where scientists submit hypotheses and plans for data analysis to a third party before performing experiments, to prevent cherry-picking statistically significant results later. Fewer than a dozen

people mentioned this strategy. One who did was Hanne Watkins, a graduate student studying moral decision-making at the University of Melbourne in Australia. Going back to her original questions after collecting data, she says, kept her from going down a rabbit hole. And the process, although time consuming, was no more arduous than getting ethical approval or formatting survey questions. "If it's built in right from the start," she says, "it's just part of the routine of doing a study."

THE CAUSE

The survey asked scientists what led to problems in reproducibility. More than 60% of respondents said that each of two factors — pressure to publish and selective reporting — always or often contributed. More than half pointed to insufficient replication in the lab, poor oversight or low statistical power. A smaller proportion pointed to obstacles such as variability in reagents or the use of specialized techniques that are difficult to repeat.

But all these factors are exacerbated by common forces, says Judith Kimble, a developmental biologist at the University of Wisconsin–Madison: competition for grants and positions, and a growing burden of bureaucracy that takes away from time spent doing and designing research. "Everyone is stretched thinner these days," she says. And the cost extends beyond any particular research project. If graduate students train in labs where senior members have little time for their juniors, they may go on to establish their own labs without having a model of how training and mentoring should work. "They will go off and make it worse," Kimble says.

WHAT CAN BE DONE?

Respondents were asked to rate 11 different approaches to improving reproducibility in science, and all got ringing endorsements. Nearly 90% — more than 1,000 people — ticked "More robust experimental design" "better statistics" and "better mentorship". Those ranked higher than the option of providing incentives (such as funding or credit towards tenure) for reproducibility-enhancing practices. But even the lowest-ranked item — journal checklists — won a whopping 69% endorsement.

The survey — which was e-mailed to *Nature* readers and advertised on affiliated websites and social-media outlets as being 'about reproducibility' — probably selected for respondents who are more receptive to and aware of concerns about reproducibility. Nevertheless, the results suggest that journals, funders and research institutions that advance policies to address the issue would probably find cooperation, says John Ioannidis, who studies scientific robustness at Stanford University in California. "People would probably welcome such initiatives." About 80% of respondents thought that funders and publishers should do more to improve reproducibility.

"It's healthy that people are aware of the issues and open to a range of straightforward ways to improve them," says Munafo. And given that these ideas are being widely discussed, even in mainstream media, tackling the initiative now may be crucial. "If we don't act on this, then the moment will pass, and people will get tired of being told that they need to do something." [SEE EDITORIAL P.437](#)

Monya Baker writes and edits for *Nature* from San Francisco.
Dan Penny aided in creation and analysis of the survey.

1. Open Science Collaboration *Science* <http://dx.doi.org/10.1126/science.aac4716> (2015).
2. Begley, C. G. & Ellis, L. M. *Nature* **483**, 531–533 (2012).
3. Patel, R. & Alahmad, A. J. *Fluids Barriers CNS* **13**, 6 (2016).
4. da Silva, C. F. et al. *Antimicrob. Agents Chemother.* **57**, 5307–5314 (2013).



ETHEL DAVIES/ROBERTHARDING/GETTY

THE SECRET HISTORY OF ANCIENT TOILETS

By scouring the remains of early loos and sewers, archaeologists are finding clues to what life was like in the Roman world and in other civilizations.

BY CHELSEA WALD

Some 2,000 years ago, a high-ceilinged room under one of Rome's most opulent palaces was a busy, smelly space. Inside the damp chamber, a bench, perforated by about 50 holes the size of dinner plates, ran along the walls. It may have supported the bottoms of some of the lowest members of Roman society.

Today, the room is shut off to the public, but archaeologists Ann Koloski-Ostrow and Gemma Jansen had a rare chance to study the ancient communal toilet on the Palatine Hill in 2014. They measured the heights of the benches' stone base (a comfortable

43 centimetres), the distances between the holes (an intimate 56 cm), the drop down into the sewer below (a substantial 380 cm at its deepest). They speculated about the mysterious source of the water that would have flushed the sewer (perhaps some nearby baths). Graffiti outside the entryway suggested long queues, in which people had enough time to write or carve their messages before taking a turn on the bench. The underground location, combined with the plain red-and-white colour scheme on the walls, implied a lower class of user, possibly slaves.

In 1913, when Italian excavator Giacomo Boni excavated this room, toilets were an

unmentionable topic. In his report, he seems to mistake the remains of the holey benches for something much more sensational: part of an elaborate mechanism that, he speculated, would have pumped water and provided power for the palace above. Boni's prudish sensibilities wouldn't let him recognize what was before his very eyes, says Jansen. "He couldn't imagine it was a toilet."

A century later, toilets are no longer such an unacceptable research topic. Koloski-Ostrow, at Brandeis University in Waltham, Massachusetts, and Jansen, an independent archaeologist based in the Netherlands, are

An ancient Roman public latrine in the ruins of Timgad, Algeria.

among a growing number of archaeologists, infectious-disease specialists and other experts who are shining light on the lost loos of history, from ancient Mesopotamia to the Middle Ages, with a particular focus on the Roman world.

Their investigations have provided a new way to learn about the diets, diseases and habits of past populations, especially those of the lower classes, which have received scant attention from archaeologists. Researchers have inferred that Roman residents ventured into their toilets with some trepidation, in part because of superstition and also because of very real dangers from rats and other vermin lurking in the sewers. And although ancient Rome is famous for its sophisticated plumbing systems, modern studies of old excrement suggest that its sanitation technologies were not doing much for the residents' health.

"Toilets have a lot to tell us about — far more than how and where people went to the bathroom," says Hendrik Dey, an archaeologist at Hunter College in New York.

QUEEN OF LATRINES

Although studies of ancient latrines are no longer off limits, they do take a certain amount of fortitude. "You have to have a very strong sense of self and of humour to work on this topic because one who works on it is going to get ribbed by friends and enemies," says Koloski-Ostrow. She got started on the topic nearly a quarter of a century ago, when classicist Nicholas Horsfall called her over in the library at the American Academy in Rome. "Latrines. Roman latrines," he whispered conspiratorially. "No one has done them properly." She took up that challenge, and now, she says, "I am known widely on my campus as 'the queen of latrines'."

The invention of some of the first simple toilets is credited to Mesopotamia in the late fourth millennium BC¹. These non-flushing affairs were pits about 4.5 metres deep, lined with a stack of hollow ceramic cylinders about 1 metre in diameter. Users would have sat or squatted over the toilet, and the excrement would have stayed inside the cylinders with the liquids seeping outwards through perforations in the rings.

Until recently, scholars had little interest in these toilets, says archaeologist Augusta McMahon at the University of Cambridge, UK. "Archaeologists in Mesopotamia have looked at them like, 'this is a problem: it's a pit that's cut into the stuff I'm really interested in.'" As far as she knows, no one has carefully excavated a Mesopotamian toilet yet — something she's hoping to do when she finds a good candidate and funding.

Mesopotamians themselves also seemed to show little enthusiasm for this revolutionary technology. Although the toilets would have been convenient to use, and cheap and easy to

install, they were uncommon, says McMahon, who surveyed the number of latrines in different neighbourhoods for a chapter in a book published last year¹. "The number of houses that have toilets is very, very low — one out of five or two out of five," she says. Everyone else probably used a chamber pot or simply squatted in the fields.

So the health benefits of the technology would have been limited, McMahon says. Although the pit toilets would have successfully separated people from their waste — the measure of a good sanitation system because it prevents the faecal–oral spread of disease — studies by the US Agency for International Development say that some 75% of a population must have access before there are widespread improvements in health.

About 1,000 years later, the Minoans on the island of Crete in the Mediterranean improved the toilet by adding the capacity to flush — although only for the elite. The first known example² was in the palace at Knossos, says Georgios Antoniou, a Greek architect who has studied ancient sanitation in that country. Water was used to wash the waste from the toilet into the sewer system of the palace.

From there, toilet technology took off. In the first millennium BC, ancient Greeks of the Classical period and, especially, the succeeding Hellenistic period developed large-scale public latrines — basically large rooms with bench seats connected to drainage systems — and put toilets into ordinary middle-class houses. "The society had become more prosperous, and they were dealing more with comfort in everyday living," Antoniou says.

"YOU HAVE TO HAVE A VERY STRONG SENSE OF SELF AND OF HUMOUR TO WORK ON THIS TOPIC."

The Romans were unprecedented in their adoption of toilets. Around the first century BC, public latrines became a major feature of Roman infrastructure, much like bathhouses, says Koloski-Ostrow. And nearly all city dwellers had access to private toilets in their residences. Nonetheless, archaeologists know very little about how these toilets worked and what people thought of them, she says. One reason is that in Roman times, few people wrote about toilets, and when they did, they were often satirical, making it hard to interpret their meaning.

But Koloski-Ostrow and Jansen show that

it is worthwhile taking the topic seriously. For a forthcoming book on toilets in the Roman capital, they and some two dozen other archaeologists have analysed more than 60 toilets scattered throughout the city, most of which had not been described before. That includes toilets for guards in the city wall, and a two-person toilet in an apartment block. "I guess it will be news to a lot of archaeologists who have worked on all kinds of Roman buildings that some of these buildings actually had toilet facilities," Koloski-Ostrow says.

Roman public latrines looked much like their Greek predecessors: rooms lined with stone or wooden bench seats positioned over a sewer. The toilet holes are round on top of the bench, and a narrower slit extends forward and down over the edge in a keyhole shape. These slits probably allowed users to insert a sponge-tipped stick for cleaning. Small gutters often run parallel to the seats along the ground; researchers suspect that people probably washed the sponges in water running through those gutters. There are no signs of barriers between the toilet seats, but people probably had a measure of privacy thanks to their long garments and the limited windows, says Koloski-Ostrow.

Private toilets were different, Jansen says. In residences, commodes were often in or near kitchens, which was practical because they were also used to dispose of food scraps. Although people flushed the toilets with buckets of water, the loos were rarely connected to sewers. When the pits filled up, they were probably emptied, either into gardens or fields outside the town, Jansen says.

Sewers — long thought to be a crowning achievement of Roman civilization — were in fact less widespread than once thought and might not have been very effective, says Koloski-Ostrow. In a book published last year³, she considered whether Roman sewers would have adhered to any of the modern principles of sanitation engineering, including regular aeration and features to control the deposition of solid waste, which would reduce the stench as well as improve flow. To a great extent, the sewers didn't meet the standards. Her own recent explorations of the Cloaca Maxima, the great sewer under Rome, revealed that some channels could get completely blocked with silt in less than a year. At the very least, they would have required regular cleaning — dirty and dangerous work.

And Roman toilets also had a number of deficiencies. One major problem was that there were no traps — or S-shaped bends — in the pipes beneath toilets to keep out flies. Environmental archaeologists Mark Robinson at the University of Oxford and Erica Rowan, now at the University of Exeter, UK, analysed the well-preserved contents of a closed sewer that was connected to several toilets in an apartment block in Herculaneum, a Roman city destroyed by an eruption of Mount Vesuvius. Among the faecal matter and other rubbish thrown down there, Robinson found lots of fragile mineralized



RIGHT: IMAGEBROKER/REX/SHUTTERSTOCK; LEFT: LUIGI SPINA/ELECTA/MONDADORI/GETTY

Communal toilets at the Roman site, Ostia Antica (left). The goddess Fortuna (right) was believed to protect latrine users from dangers.

fly pupae. With easy access to human waste, flies could have transferred faecal matter and pathogens to people.

To look at the benefits of ancient sanitation systems, palaeopathologist Piers Mitchell at the University of Cambridge analysed published studies of parasites found at archaeological sites from several eras⁴. Contrary to his expectations, the prevalence of intestinal parasites such as roundworm and whipworm — which cause dysentery — did not decrease from the Bronze and Iron ages to the Roman period; they gradually rose. That might be because the Romans used human waste as fertilizer, which would have transferred the parasite eggs to food. “Toilets and sewers and things didn’t seem to improve the intestinal health of the Roman population,” he says.

DIET DETAILS

The practice of throwing kitchen rubbish down toilets was unhygienic for the ancient Romans, but the remnants of that refuse are now a rich source of information. Rowan was surprised by the quality and variability of the foods in the Herculaneum sewer, especially because it was connected to an apartment complex that housed a large number of mostly poorer people. “We always think that anyone non-elite in the ancient world is not eating a very diverse or interesting diet,” she says. But the evidence from Herculaneum shows that people across the class spectrum were eating dozens of different types of food, most commonly figs, eggs, olives, grapes and shellfish. They flavoured their meals with seasonings such as dill, mint, coriander and mustard seeds⁵. “It would be quite healthy, and they’d be getting all their essential nutrients.”

Rowan also used the sewer contents to glean insights into the broader food and energy economy. The large amount of kitchen scraps suggested that the residents cooked more at home than previously thought⁵. From the quantity of fish bones found, she concluded that the regional fish trade was

probably much larger than scholars had suspected⁶. Such discoveries are part of a broader trend in Roman archaeology, says Dey. Until recently, most scholars focused on the monumental structures occupied by elite residents. But attention has shifted to lower down the class ranking. “Roman archaeologists started to realize that you can’t understand how a society works if you only study the 1%,” he says. “The study of toilets is part of the broader effort to understand how Roman society worked, which includes — especially — studying how the non-glamorous parts of society worked.”

For Koloski-Ostrow and Jansen, latrines provide a window onto the beliefs of that society. Romans perceived demons everywhere, and some Roman literature refers to ones that lurked in toilets. “The demons can cast a spell on you, and when you have this spell you die or you get sick,” Jansen says.

The Roman writer Claudius Aelianus tells a story in his *De Natura Animalium* about an octopus that swam up through a drain in a toilet and ate the pickled fish in the pantry night after night. That story is probably apocryphal, but rodents, insects and other creatures could have lurked in toilets and invaded homes. And excrement-filled water could have flowed upwards during flooding.

Explosive gases might also have been a problem. “You might walk in and actually see a flame burst out of one of those holes because of the methanolic gases that built up in the sewer underneath the toilet,” Koloski-Ostrow speculates.

This pervasive fear of toilets could explain the mystery of why there’s less graffiti inside public latrines than in the rest of the Roman world, Jansen says. Nobody wanted to spend more time there than necessary. The same fear could also explain why many latrines have small shrines to the goddess Fortuna. Jansen argues that she was thought to protect toilet-users from illness-causing demons, as well as the other bad things that could happen there⁷.

More discoveries about ancient lifestyles will come as researchers expand their toilet studies to other parts of the globe. Rowan is studying a site in Turkey, and Mitchell has recently examined evidence from a 2,000-year-old toilet in China. But progress has been slow and archaeologists are not rushing into toilet studies. Although the topic is no longer considered fringe, funding is hard to come by, and Mitchell says that “no one else seems to be that bothered” to work on it. One reason could be that the lack of written sources and the limited physical evidence make it daunting.

But for researchers such as Koloski-Ostrow, the recent work raises all kinds of questions about ancient societies. Did women use public toilets? Were they chatty, social places or silent? What were the foreign influences on Roman toilets, and how did the toilet culture propagate between the capital and the distant states? These questions will be hard to answer, she says, but asking them no longer seems as weird as when she started.

Rowan agrees: toilets have finally gone mainstream. “If somebody finds a latrine now, they know to sample it, to excavate it carefully. They know there’s going to be a lot of value in it, as opposed to being, like, oh, it’s just a toilet.” ■

Chelsea Wald is a journalist in Vienna, Austria.

1. McMahon, A. in *Sanitation, Latrines and Intestinal Parasites in Past Populations* (ed. Mitchell, P. D.) 19–40 (Routledge, 2015).
2. Antoniou, G. P. & Angelakis, A. N. in *Sanitation, Latrines and Intestinal Parasites in Past Populations* (ed. Mitchell, P. D.) 41–68 (Routledge, 2015).
3. Koloski-Ostrow, A. O. *The Archaeology of Sanitation in Roman Italy* (Univ. North Carolina Press, 2015).
4. Mitchell, P. D. *Parasitology* <http://dx.doi.org/10.1017/S0031182015001651> (2016).
5. Robinson, M. & Rowan, E. in *A Companion to Food in the Ancient World* (eds Wilkins, J. & Nadeau, R.) 105–115 (Wiley-Blackwell, 2015).
6. Rowan, E. in *Fish & Ships: Production et Commerce des Salsamenta Durant l’Antiquité* (eds Botte, E. & Leitch, V.) 61–74 (Errance, 2014).
7. Jansen, G. C. M. et al. (eds) *Roman Toilets: Their Archaeology and Cultural History*. BABESCH Suppl. 19 (Peeters, 2011).

COMMENT

PHYSICS Which next-generation neutrino facility should be funded? **p.462**

ARCHAEOLOGY Sea-bed trove from Ancient Egypt's long-sunk cities on show **p.466**

HEALTH A memoir of a life leading the fight against smallpox and HIV **p.468**

OBITUARY Harry Kroto, buckyball co-discoverer, remembered **p.470**

UIG/GETTY



Surgery can be an effective treatment for type 2 diabetes.

Time to think differently about diabetes

New guidelines for the surgical treatment of type 2 diabetes bolster hopes of finding a cure, writes **Francesco Rubino**, but long-standing preconceptions must be put aside.

Clinical guidelines published this week¹ announce what may be the most radical change in the treatment of type 2 diabetes for almost a century. Appearing in *Diabetes Care*, a journal of the American Diabetes Association, and endorsed by 45 professional societies around the world, the guidelines propose that surgery involving the manipulation of the stomach or intestine be considered as

a standard treatment option for appropriate candidates. This development follows multiple clinical trials showing that gastrointestinal surgery can improve blood-sugar levels more effectively than any lifestyle or pharmaceutical intervention, and even lead to long-term remission of the disease¹.

As someone who has been investigating the link between gastrointestinal surgery and glucose homeostasis since the late 1990s

(see 'Surgical breakthrough'), I have witnessed first-hand how getting to this point has required many clinical scientists to put aside long-standing preconceptions. Indeed, the guidelines come nearly 100 years after the first clinical observations that diabetes could be improved or even resolved by a surgical operation (see 'A long road')². The evidence that surgery can prompt the remission of a disease that has long been considered ▶

DIABETES TREATMENT

Surgical breakthrough

In 1925, a report in *The Lancet*² described a 'side effect' of a gastrointestinal operation to treat a peptic ulcer. This was the almost overnight resolution of an excess of sugar in the urine (glycosuria) — the chief symptom of diabetes at the time. Similar observations were reported in subsequent decades and became more common after the advent of bariatric or weight-loss surgery in the mid 1950s, which led to more people with diabetes receiving these types of operations. And during the 1980s and 1990s, resolution of diabetes after bariatric surgery was noted on many occasions, including in a landmark report involving more than 120 patients⁹.

In 1999, while working as a research fellow at Mount Sinai School of Medicine in New York City, I stumbled across a report showing that nearly all people with type 2 diabetes who had undergone a complex bariatric operation (biliopancreatic diversion) had completely normal blood-sugar levels as early as one month after surgery. They had been able to stop taking medication and come off a low-calorie diet. I wondered whether gastrointestinal surgery could influence diabetes directly. If so, surgery could be used to treat diabetes or to understand how it works.

The next day, I persuaded my mentor to seek approval from the institutional review board to run trials in humans. Failing to obtain approval, we turned to rats to investigate whether a modified form of gastric-bypass surgery could directly influence glucose homeostasis. Our experiments confirmed that it could, although it took us more than two years to publish the findings⁶.

In 2006 and 2007, surgical teams showed that the operation had the same effect in humans¹⁰, and other groups began to investigate the molecular mechanisms that might be responsible. On the back of these studies, a multidisciplinary group of leading clinicians and scientists at the first Diabetes Surgery Summit in 2007 reviewed the preliminary mechanistic and clinical data available on the effects of surgery on diabetes and established an agenda for research priorities. The summit inspired the randomized clinical trials that now provide the evidence supporting a role of surgery in diabetes. In September 2015, the introduction of surgery into standard care for type 2 diabetes was formally recommended by the participants of the second Diabetes Surgery Summit¹. **F.R.**

► irreversible could bolster searches for what causes diabetes and even reinvigorate hopes to find a cure. But future progress will require more thinking outside the box.

CLINICAL SHIFT

The number of adults around the world with diabetes quadrupled from 108 million in 1980 to 422 million in 2014 (ref. 3). About 90% of these people have type 2 diabetes — a major cause of kidney failure, blindness, nerve damage, amputations, heart attack and stroke. Fewer than 50% of people with type 2 diabetes control their blood-sugar levels adequately by changing their diet or exercise regime, or by taking drugs.

Bariatric or weight-loss surgery refers to various procedures. Surgeons may, for instance, remove a portion of the person's stomach or divide the stomach into two and reroute the small intestine to the upper part (see 'Gastric bypass'). Since the mid 1950s, people whose body mass index (BMI) is greater than 40 have received bariatric surgery to induce weight loss. Many of these people also had diabetes. The new guidelines advise that such procedures (metabolic surgery) be considered specifically for the treatment of diabetes in people who have not adequately controlled their blood-sugar levels through other means, and whose BMI is greater than 30 (or 27.5 for people of Asian descent). Perhaps more significantly, they also state that the gastrointestinal tract is an appropriate biological target for interventions designed to treat diabetes¹.

These recommendations arguably signify the most radical departure from mainstream approaches to the management of diabetes since the introduction of insulin in the 1920s. They are based on findings from a large body of work, including 11 randomized clinical trials conducted over the past decade¹. In these studies, most

surgically treated people (up to 80% in a recent 5-year follow-up⁴ of a randomized trial) fall into one of two categories. Either their diabetes goes into apparent remission or their blood-sugar levels can be stabilized using reduced medication or exercise and a calorie-controlled diet (see 'Big benefits').

Non-randomized studies, involving people receiving surgery and matched subjects treated with standard interventions, suggest that surgery may also reduce heart attacks, stroke and diabetes-related mortality¹. And several economic analyses suggest that the costs of surgery (roughly US\$20,000–25,000 per procedure in the United States) may be recouped within 2 years through reduced spending on medication and care⁵.

The effects of surgery on diabetes are dramatic. Yet it has taken nearly a century to unearth them since observations of major improvement or remission of diabetes after surgical operations were first reported².

A major stumbling block seems to have been the lack of a plausible mechanism to explain how gastrointestinal surgery is able to resolve the symptoms of diabetes. Numerous surgeries — knee and hip replacements, appendix removal, even bariatric surgery — have been performed for decades without randomized trials confirming that these approaches are more effective than less invasive ones. But surgery explicitly seems to fix what is broken in those instances. However, in the case of diabetes — a systemic disease with dysfunctions involving the pancreas, liver, muscle and fat (adipose) tissue — it has been much harder to imagine what surgery would be able to mend.

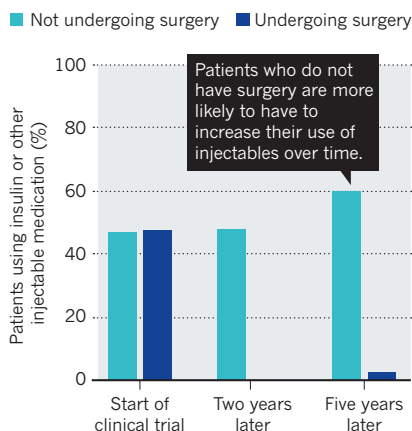
The dominant 'adipocentric model' has also been a major conceptual barrier to the acceptance of surgery as a treatment for the disease itself. This model posits that excess fat causes diabetes, either by causing the liver to malfunction or by making other cells resistant to insulin. Because this model predicts that the reduction of fatty tissue, however obtained, can relieve the symptoms of diabetes, weight loss after bariatric surgery has provided a straightforward explanation for the associated remission of the disease.

It was exactly this absence of understanding about mechanism — and the mismatch between observations and mainstream thinking — that delayed the prescription of the painkiller aspirin to people with heart disease in the twentieth century. Clinical observations in the early 1950s suggested that aspirin could prevent thromboses. But large-scale trials to test the drug's ability to prevent heart attacks began only in the 1970s, after experiments had shown that it could inhibit blood clotting.

We now know that the dramatic effects of surgery on diabetes are not just a consequence of weight loss. Changes to gastrointestinal

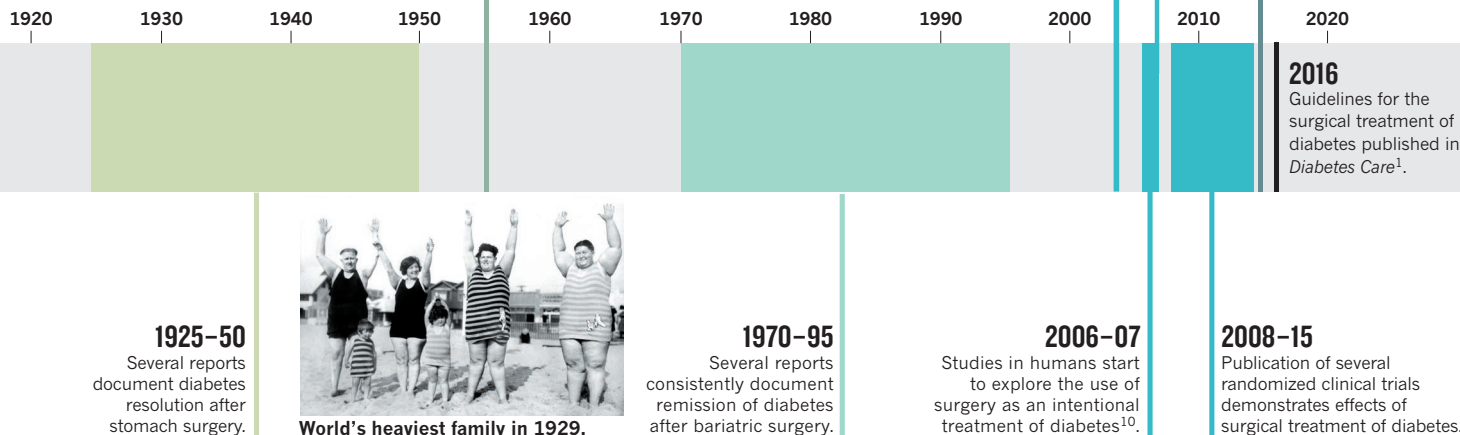
BIG BENEFITS

In one clinical trial, most people with diabetes who had received gastrointestinal surgery did not need to take insulin or other injectable medication to control their blood-sugar levels even five years after the operation.



A LONG ROAD

Observations that diabetes can be improved or even resolved by surgical operations have been reported for almost a century.



anatomy can directly influence glucose homeostasis⁶. Over the past decade, efforts to explain the link have identified several potential mechanisms⁷. For one, surgery seems to alter the amount and timing of the secretion of gut hormones, which in turn influence insulin production. Experiments also suggest that surgery can increase the production of certain bile acids that make cells more sensitive to insulin, or increase the uptake of glucose by the gut cells themselves, thereby lowering blood glucose levels. Surgery-induced changes to the composition of the gut microbiota and to the efficiency of intestinal nutrient sensing also seem to contribute. This is the process by which cells lining the gut detect certain nutrients and send neural signals to brain centres involved in the regulation of glucose metabolism.

A CHANGE OF MIND

Capitalizing on these latest insights about type 2 diabetes will require a shift in mindsets across the broad spectrum of care and research.

The high upfront costs of surgery and the specialized staff and medical centres needed to deliver it make surgery an unlikely solution for the ongoing epidemic. Rates of diabetes are rising rapidly in low- and middle-income countries³, where surgery is not likely to be available for most patients. But if handled the right way, the inclusion of surgery as an option could influence diabetes care as a whole.

Currently, many people with diabetes and obesity grow disheartened after trying one treatment after another to no avail. Just knowing that through surgery the possibility of major improvement and even remission exists may be empowering to some. Also, to identify those people for whom surgery may be appropriate, providers will first need to be confident that other options have failed¹. So

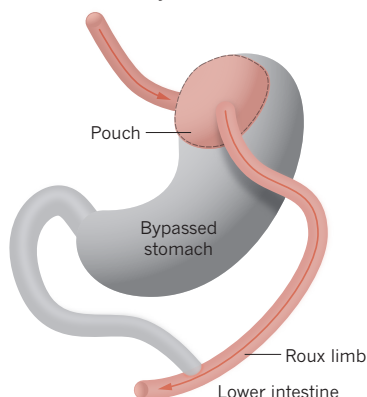
both patients and providers may be encouraged to approach conventional treatments with more determination and rigour.

The broad endorsement of surgery as a treatment option should also inspire fresh approaches in research. Researchers and clinicians are already trying to mimic the effects of gastrointestinal surgery using less-invasive interventions. For instance, experiments originally conducted in rats, and various studies in humans, have shown that blocking intestinal signalling from the duodenum or upper small intestine can alleviate the symptoms of diabetes⁶.

One approach aims to do this by means of a tube inserted into the intestine. Designed to prevent contact between nutrients and the lining of the upper small intestine, this tube mimics the effects of a surgical bypass. The device has been approved for clinical use in Europe and Australia. Another approach involves passing a balloon-tipped device through the mouth and down into the duodenum, where it is filled with hot water to burn (ablate) the cellular lining. A clinical trial of the device is currently in the

GASTRIC BYPASS

The stomach is reduced to a small pouch and is connected directly to the intestine.



recruitment stage in Europe. Pharmacological interventions that target gastrointestinal mechanisms of metabolic regulation are also being investigated.

The ability of gastrointestinal surgery to influence glucose homeostasis and clinically reverse diabetes suggests that the disease might be explained, at least in part, by a fault in the mechanisms through which the gut regulates metabolism⁸. Testing this hypothesis could provide insights about ways to prevent and even cure diabetes.

Despite the compelling results of clinical trials and experimental work on nutrient–gut signalling mechanisms, shifting diabetes' image as an incurable, hopeless condition caused by excess fat will still require some imagination. Albert Einstein once said that imagination is more important than knowledge. The story of surgery and diabetes shows how important it is to have both. ■

Francesco Rubino is professor of metabolic and bariatric surgery at King's College London and consultant surgeon at King's College Hospital, London, UK. He is an organizer of the Diabetes Surgery Summit. e-mail: francesco.rubino@kcl.ac.uk

1. Rubino, F. *et al. Diabetes Care* **39**, 861–877 (2016).
2. Leyton, O. *Lancet* **206**, 1162–1163 (1925).
3. NCD Risk Factor Collaboration *Lancet* **387**, 1513–1530 (2016).
4. Mingrone, G. *et al. Lancet* **386**, 964–973 (2015).
5. Klein, S., Ghosh, A., Cremieux, P. Y., Eapen, S. & McGavock, T. J. *Obesity* **19**, 581–587 (2011).
6. Rubino, F. & Marescaux, J. *Ann. Surg.* **239**, 1–11 (2004).
7. Dixon, J. B., Lambert, E. A. & Lambert, G. W. *Mol. Cell. Endocrinol.* **418**, 143–152 (2015).
8. Rubino, F. *Diabetes Care* **31**, S290–S296 (2008).
9. Pories, W. J. *et al. Ann. Surg.* **222**, 339–350 (1995).
10. Cohen, R. V., Schiavon, C. A., Pinheiro, J. S., Correa, J. L. & Rubino, F. *Surg. Obes. Relat. Dis.* **3**, 195–197 (2007).

The author declares competing financial interests: see go.nature.com/49vyhn for details.



An optical sensor begins its 2,500-metre journey down a borehole to become part of the IceCube neutrino detector in Antarctica.

Invest in neutrino astronomy

Spencer Klein calls for bigger telescope arrays to catch particles from the most energetic places in the Universe.

Neutrino astronomy is poised for breakthroughs. Since 2010, the IceCube experiment in Antarctica — 5,160 basketball-sized optical sensors spread through a cubic kilometre of ice — has detected a few score energetic neutrinos from deep space. Although these are exciting finds that raise many questions, this paltry number of extraterrestrial particles is too few to tell their origins or to test fundamental physics. To learn more will require a new generation of neutrino observatories.

Neutrinos are subatomic particles that

interact only weakly, so they can travel far through space and even penetrate Earth. IceCube detects highly energetic neutrinos, with energies above about 100 gigaelectronvolts (1 GeV is 10^9 electronvolts, roughly the rest mass of a proton). These are produced when cosmic rays — high-energy protons or heavier nuclei from space — interact with matter or light. This might happen either at the sites where the cosmic rays are produced, or later when the rays enter Earth's atmosphere and collide with gas molecules, releasing a cascade of elementary particles.

Neutrinos produced in the atmosphere are hundreds of times more numerous than the astrophysical ones.

Many physics puzzles stand to be solved by neutrino astronomy¹. One is the origin of the ultra-high-energy cosmic rays. In 1962, the Volcano Ranch array in New Mexico detected an enormous shower of particles coming from one cosmic ray smashing into the upper atmosphere with a kinetic energy of above 10^{11} GeV — equivalent to the energy of a tennis serve packed into a single atomic nucleus. Tens more such events have been

BLAINE GUDBJARTSSON, ICECUBE/NSF

detected since. But 50 years on, physicists still have no idea how nature accelerates elementary particles to such high energies. The energies far exceed the range of Earth-bound accelerators such as the Large Hadron Collider (LHC) near Geneva, Switzerland; mimicking them would require a ring the size of Earth's orbit around the Sun.

There is also much we need to find out about neutrinos themselves — their accurate masses, how they transform from one type (flavour) into another, and whether other predicted forms (such as 'sterile' neutrinos) exist. Neutrinos could also help to find dark matter, invisible material that has a part in controlling the motions of stars, gas and galaxies. Decaying or annihilating dark matter could produce energetic neutrinos that would be visible to neutrino telescopes.

The downside of neutrinos' weak interactions is that an enormous detector is required to catch enough particles to distinguish the few space-borne ones from the many more originating from Earth's atmosphere. IceCube is the largest neutrino-detection array in operation but it is too small, and further data collection is probably too slow to yield major breakthroughs in the next decade.

Bigger neutrino observatories, with volumes that are 10–100 times greater than that of IceCube, are essential to explore the most energetic processes in the Universe. Determining the masses of different types of neutrino and studying how neutrinos interact with matter within Earth could distinguish or rule out some models of extra spatial dimensions and address key concerns for high-energy nuclear physics such as the density of gluons (which mediate forces between quarks) in heavy nuclei.

Designs for neutrino telescopes are on the drawing board and could be up and running in five to ten years — if the astro-, particle- and nuclear-physics communities can come together and coordinate funding. A complementary set of several neutrino observatories would test physics at energies beyond the LHC's at a fraction of the cost — tens to hundreds of millions, rather than tens of billions, of dollars.

MORE QUESTIONS THAN ANSWERS

IceCube, which became fully operational in Antarctica in 2010 (and with which I have been involved since 2004), detects blue light: Cherenkov radiation that is emitted by the charged particles produced when energetic neutrinos interact with atomic nuclei in water or ice. Computers comb through the data to look for interactions — long tracks or radial cascades of particles emanating from a point (see 'Neutrino observatory'). IceCube sees more than 50,000 neutrino candidates per year. Fewer than 1% are from space.

There are several ways to distinguish cosmic from atmospheric neutrinos. The

NEXT-GENERATION NEUTRINO TELESCOPES

Bigger neutrino arrays have been proposed that would catch enough cosmic neutrinos to probe extreme energies and test basic physics and astronomy.

Experiment	Detects	Where	Volume	Estimated cost (US\$)
IceCube-Gen2	Optical	South Pole, Antarctica	10 km ³	\$400 million
Cubic Kilometre Neutrino Telescope (KM3NeT)	Optical	Mediterranean Sea (two sites being considered)	5 km ³	\$250 million
Gigaton Volume Detector	Optical	Lake Baikal, Russia	1 km ³	Unknown
Askaryan Radio Array	Radio	South Pole, Antarctica	>100 km ³	\$5 million to \$30 million
ARIANNA	Radio	Ross Ice Shelf, Antarctica	>100 km ³	\$5 million to \$30 million

highest-energy events are more likely to be astrophysical. Atmospheric neutrinos are accompanied by particle showers, which can be seen with detectors on the ice surface. Muons, short-lived subatomic particles produced in these showers, are 500,000 times more numerous than neutrinos, and can also penetrate the ice; so signals accompanied by muons travelling downwards from the sky are probably atmospheric in origin. That leaves extremely energetic events with light trails that are travelling upwards (through Earth) or that originate from a point within the array volume as potentially astrophysical in origin.

Since 2010, IceCube has seen about 60 astrophysical neutrino candidates^{2,3}. Other experiments are too small to detect any such neutrinos; these include ANTARES, an array of strands of detectors anchored to the floor of the Mediterranean Sea off Marseilles, France, and another similar array in Lake Baikal, Russia. Their detection rate of astrophysical neutrinos is as high as could be expected — if there were more neutrinos, they would drain the cosmic rays of most of their energy⁴. So finding the astrophysical sources of the neutrinos should be easy. The fact that we have not is a growing puzzle.

So far, neutrinos do not seem to be coming from particular sites on the sky⁵, although several groups have suggested a weak link to the plane of the Milky Way. And analyses disfavour the many sites once thought likely to have accelerated energetic cosmic rays and neutrinos, including γ -ray bursts (GRBs) and active galactic nuclei (AGNs).

GRBs are short bursts of powerful γ -rays that are picked up by satellites. They are

thought to emanate either from a black hole coalescing with a neutron star or another black hole (producing a rapid burst last-

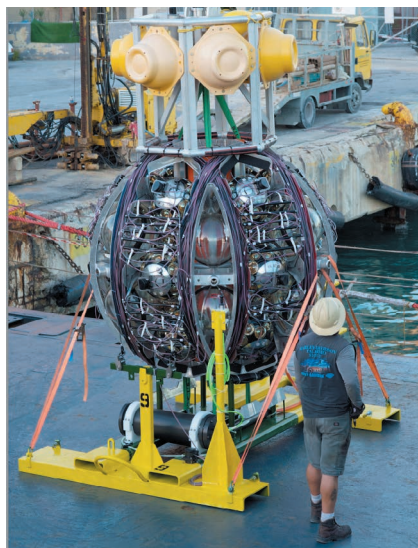
ing less than 2 seconds); or from the slower collapse of supermassive stars (bursts lasting seconds or minutes). Particles are accelerated by the implosion or explosion. Of more than 800 GRBs examined by IceCube scientists, none was accompanied by a burst of neutrinos, implying that GRBs can produce at most 1% of the astrophysical neutrinos seen by IceCube⁶.

AGNs are galaxies that at their centres have supermassive black holes accreting gas. Particles may be accelerated to relativistic speeds in jets of material that are blasted out from the black hole. But IceCube sees no associations between energetic neutrinos and active galaxies with jets that point towards Earth, suggesting that active galaxies explain at most 30% of the neutrinos⁷.

Other unlikely sources include starburst galaxies, which contain dusty regions of intense star formation that are riddled by supernova explosions⁸; magnetars, which are neutron stars surrounded by strong magnetic fields that expel powerful bursts of neutrinos for a few days (these should have been seen by IceCube); and supernova remnants, whose magnetic fields are too weak to explain the most energetic neutrinos⁹, even though they are believed to be responsible for most lower-energy (up to about 10^{16} eV) cosmic rays seen in the Galaxy.

More exotic possibilities remain untested: as-yet-unseen supermassive dark-matter

“Finding the astrophysical sources of the neutrinos should be easy.”



A string of optical modules of the KM3NeT array.

particles that annihilate and produce energetic neutrinos; or the decay of cosmic 'strings', discontinuities in space-time left over from the Big Bang.

IceCube has also tested alternative physics theories. It has constrained how neutrinos 'oscillate' from one flavour to another and set limits on the properties of dark matter and on the constituents of high-energy air showers.

NEXT GENERATION

There are two ways forward: enlarge the current optical arrays to collect more neutrinos, or find other strategies for isolating the highest energy neutrinos that must be cosmic in origin. These approaches cover different energy ranges and thus complementary physics. Both merit support.

First, larger optical Cherenkov telescopes could be deployed in ice or a lake, sea or ocean — similar to IceCube or ANTARES but with more efficient optical sensors and cheaper technology. Several groups have developed advanced designs for these concepts but lack funding. The detectors could be constructed and operational by the early 2020s. For IceCube, technical improvements would include more efficient drilling technology and sensors that fit in narrower bore holes, which are cheaper to drill.

Different sites offer different benefits. Antarctica offers a large expanse of clear, compacted ice and infrastructure. But arrays in the Northern Hemisphere, for example, in the Mediterranean, can more directly observe astrophysical neutrinos from the centre of the Galaxy that have passed through Earth, without having to reject down-going atmospheric neutrinos, as a southern site would have to. The absence of potassium-40 and the lower bioluminescence in fresh water (which contribute to background light and can confuse the reconstruction of particle tracks), and the presence of a frozen surface during the winter, simplifying construction, make Lake Baikal an attractive site.

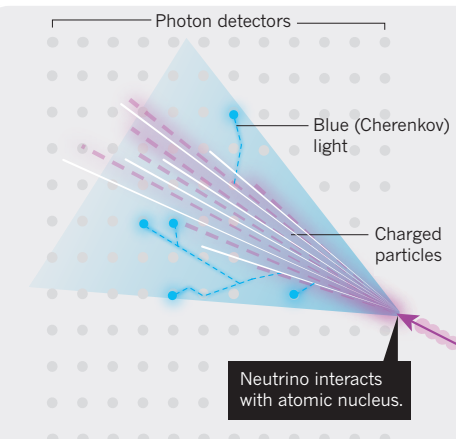
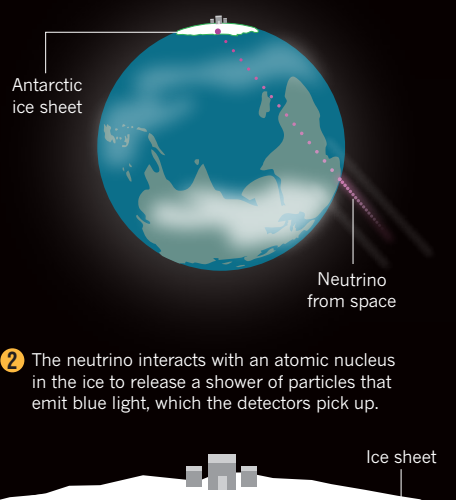
The second approach requires catching neutrinos with energies above 10^8 GeV. Neutrinos this energetic are rare — IceCube has seen none — and an array of at least 100 km^3 would be needed to capture enough events. Because optical Cherenkov light travels only tens of metres in ice or water, covering such a volume would require millions of sensors and would be expensive.

A more practical way is to search for radio emissions from neutrino interactions with the Antarctic ice sheet. When the neutrinos hit an atomic nucleus in the ice, they create a shower of charged particles that give off radio waves in the 50 megahertz to 1 gigahertz frequency range, as well as visible light. Radio waves can propagate for kilometres through ice. So an radio-sensing array over 100 km^3 could be more sparsely populated with instruments, with roughly one station per cubic

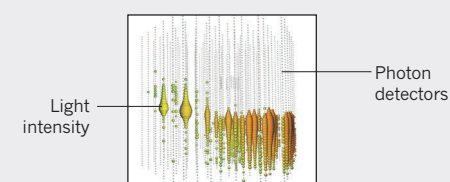
NEUTRINO OBSERVATORY

As a neutrino from space (1) interacts with atomic nuclei in water or ice, the shower of particles given off emit blue light (2). The light can be tracked with detectors to reveal the neutrino's energy and the direction the particle came from (3).

1 Cosmic neutrino passes through Earth.



3 Strings of buried optical sensors track blue light from the particles as they move through the array.



kilometre. The radio pulses from neutrinos with energies above 10^8 GeV should be strong enough for antennas in the ice to pick up. Two international groups are building prototypes and have sought funding to expand (I am involved with one, ARIANNA).

GREEN LIGHT

With a range of affordable, next-generation designs shovel-ready, decisions about design priorities need to be made and

grants deployed. The main obstacles are limited national science budgets and funding-agency silos. Neutrino astronomy falls between the particle-, nuclear- and astrophysics communities, which need to pool resources to realize the promise of these techniques.

First, one or both of the successors to IceCube and ANTARES should be funded and built. An upgraded IceCube experiment (IceCube-Gen2) and the Cubic Kilometre Neutrino Telescope (KM3NeT), a proposed European project, are both strong candidates (see 'Next-generation neutrino telescopes'). If necessary, the teams coordinating IceCube, KM3NeT and the Gigaton Volume Detector¹⁰, a proposed Russian array, should explore merging these collaborations to focus on a single large detector at the most cost-effective site. Funding should be sought from a wider range of agencies, including those focused on particle and nuclear physics.

Second, at least one 100-km^3 radio-detection array needs to get the go ahead. Because such a project can be done only in Antarctica, the onus is on the US National Science Foundation, which is the largest supporter of Antarctic research and realistically the only group that has adequate logistical resources to pull off such a project. Many non-US groups are interested, and collaborations should be set up and costs shared internationally. Once proven, such an array could be expanded to cover $1,000 \text{ km}^3$ by around 2030 to monitor the ultra-high-energy Universe.

By finding the astrophysical sources of ultra-energetic neutrinos and cosmic rays — or ruling out remaining models — the next generation of neutrino observatories is guaranteed to make discoveries. ■

Spencer Klein is a senior scientist in the Nuclear Science Division, Lawrence Berkeley National Laboratory, and a research physicist at the University of California, Berkeley, Berkeley, California, USA.
e-mail: srklein@lbl.gov

- Halzen, F. & Klein, S. R. *Phys. Today* **61N5**, 29–35 (2008).
- Aartsen, M. G. *et al. Phys. Rev. Lett.* **111**, 021103 (2013).
- IceCube Collaboration. Preprint at <https://arxiv.org/abs/1510.05223> (2015).
- Bahcall, J. & Waxman, E. *Phys. Rev. D* **64**, 023002 (2001).
- IceCube Collaboration. Preprint at <https://arxiv.org/abs/1510.05222> (2015).
- Aartsen, M. G. *et al. Astrophys. J.* **805**, L5–L12 (2015).
- DeYoung, T. *EPJ Web Conf.* **116**, 11004 (2016).
- Bechtol, K. *et al.* Preprint at <https://arxiv.org/abs/1511.00688> (2015).
- Chakraborty, S. & Izaguirre, I. *Phys. Lett. B* **745**, 35–39 (2015).
- Avrorin, A. D. *et al.* Preprint at <http://arxiv.org/abs/1511.02324> (2015).



A colossal statue of Hapy, the ancient Egyptian god of Nile flooding, being raised from Abu Qir Bay.

Sunken Cities: Egypt's Lost Worlds

British Museum,
London.
Until 27 November
2016.

operation by the European Institute for Underwater Archaeology (IEASM) in Paris, directed by Franck Goddio. The team used side-scan sonar,

nuclear magnetic resonance magnetometers and sub-bottom profilers to reveal slices through the geological strata beneath the sea bed, allowing them to begin partial excavation. Divers uncovered buildings, massive sculptures and a huge range of objects, from bronze incense burners to gold jewellery. More than 750 ancient anchors and 69 ships were also detected in the ooze, most of them from the sixth to second centuries BC, in Thonis-Heracleion's harbour.

An exhibition of this extraordinary trove, *Sunken Cities*, is the first large-scale show of underwater discoveries at the British Museum in London, and the most complete presentation of this complex Egyptian-Greek society so far.

More than 200 IEASM finds are exhibited, denoted on the information panels by a hieroglyphic zigzag symbolizing water. Many of them were displayed in Berlin's Martin-Gropius-Bau and the Grand Palais in Paris in 2006–07, but much has been discovered since. Among the stone statues of deities and rulers in pharaonic or Greek dress is a 5.4-metre figure of Hapy, god of the Nile inundation, which greets visitors as it once greeted Greek sailors approaching the mouth of the Nile. Nearby are inscriptions in hieroglyphic and Greek on stone and gold, intricate jewellery in recognizably Greek styles and delicate lead models of votive barques used in the cult of Osiris. These exhibits are supplemented by objects from other sites, lent by a number of museums in Alexandria and by the Egyptian Museum in Cairo, as well as the British Museum's renowned Egyptian collections (notably, those from the upstream ancient Greek port of Naukratis). Ethereal silent underwater film footage strategically positioned throughout the exhibition shows divers investigating and rescuing a few key objects, including a sycamore barge of Osiris.

The hybrid culture on show may surprise, and at times confuse, visitors familiar with the art and objects characteristic of earlier Egyptian dynasties found at sites such as Luxor and Abu Simbel. Says exhibition curator Aurélia Masson-Berghoff: "People sometimes assume that when two cultures mix, the essence of each is diluted and, as a result, weakened; *Sunken Cities* demonstrates the opposite." She notes that ancient Egypt was not isolated, as is sometimes thought, but an "outward looking, influential and inclusive" society. This is amply borne out in the history and culture of Ptolemaic and Roman Egypt. Alexander and his friend and general Ptolemy — who became Ptolemy I

FRANCK GODDIO/HILTI FOUNDATION/CHRISTOPH GERIGK

ARCHAEOLOGY

Soaked in history

Andrew Robinson tours an enthralling exhibition of finds from two ancient cities, long sunk in the Nile delta.

In Egypt's Nile delta at Abu Qir Bay lie the remains of two cities. In 500 BC, Canopus and Thonis-Heracleion were crucial ports for trade with Greece and Greek settlement. Even their names reflect Greek mythological figures: Kanopos, pilot of Spartan king Menelaus's ship in the Trojan War, and the hero Herakles. The cities continued to flourish until at least the late fourth century AD, through Alexander the Great's founding of Alexandria in 332 BC, the Ptolemaic

period of Greek rule that ended in 30 BC, and Roman rule.

At some point, however, they began to sink. By the eighth century AD, the cities were submerged several metres under the Mediterranean sea bed, their precise locations lost for centuries.

➔ NATURE.COM
For more on science
in culture see:
[nature.com/
booksandarts](http://nature.com/booksandarts)

In the late 1990s, they were found as part of a technically challenging and scientifically sophisticated

in 306 BC — worshipped at Egyptian shrines; and Cleopatra and her lover the Roman general Mark Antony presented themselves as the living Egyptian–Greek deities Isis–Aphrodite and Osiris–Dionysus. A terracotta lamp with an Egyptian Isis motif, dating from the second century AD and on show from the British Museum collection, was found in far-off Roman Britain.

Ancient Egyptian science also appears. The black granite of a fascinating shrine from the fourth century BC known as the Naos of the Decades is heavily inscribed with hieroglyphs and a large figure of a lion. In submerged Canopus, it broke apart and the pieces spread far and wide: the roof ended up in the Louvre in Paris in 1817; the base and rear wall were found on site underwater in 1933 and deposited in the Graeco-Roman

“Ancient Egypt was not isolated, but an outward looking, influential and inclusive society.”

Museum in Alexandria. Amazingly, the IEASM team stumbled on four further pieces in 1999. Egyptologist Anne-Sophie von Bomhard examined the Naos, reconstructed after

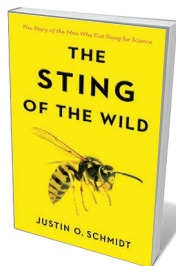
more than a millennium. Its external surfaces depict a calendar dividing the Egyptian year into ten-day sections (‘decades’), connected with the successive rising of certain stars (‘decans’). This proved that ancient Egyptian astrology was based on astronomical observations.

The exhibition discusses theories about why Canopus and Thonis-Heracleion sank without favouring any one cause, for lack of definitive historical or contemporary evidence. Possibilities range from tsunamis and earthquakes to floods, variations in sea level and geological subsidence, all known to have occurred in the region. There was, for instance, an earthquake in AD 796 or 797 that damaged the top section of Alexandria’s Pharos lighthouse, according to ninth-century AD Arab historian al-Tabari. These natural forces may have contributed to another, human-made, phenomenon revealed by core samples taken from the sediments under Abu Qir Bay: the liquefaction of the clay soils, triggered by pressure from the cities’ heavy temple buildings.

According to Goddio, perhaps as little as 5% of the area around the sunken cities has been investigated. As he writes at the exhibition’s end: “What we know now is just a fraction. We are still at the very beginning of our search.” ■

Andrew Robinson is the author of *Cracking the Egyptian Code: The Revolutionary Life of Jean-François Champollion*.
e-mail: andrew@andrew-robinson.org

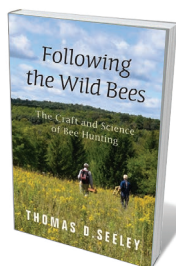
Books in brief



The Sting of the Wild

Justin O. Schmidt JOHNS HOPKINS UNIVERSITY PRESS (2016)

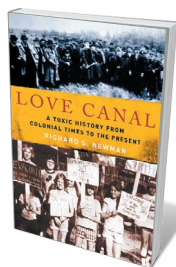
Entomologist Justin Schmidt takes an immersive approach to his work. Notching up the stings of 83 insect species, he ranks them on a pain index from ‘ethereal’ to ‘satanic’. His low-down on sting biochemistry and physiology is relentlessly zestful, even as he recounts the swelling, burning consequences of his curiosity. We also meet perpetrators such as the bullet ant (*Paraponera clavata*), equipped with a fearsome sting and chemical warnings smelling of burnt garlic; and the tarantula hawk wasp (*Pepsis* spp.), whose scream-inducing jab delivers mysteriously non-toxic venom.



Following the Wild Bees: The Craft and Science of Bee Hunting

Thomas D. Seeley PRINCETON UNIVERSITY PRESS (2016)

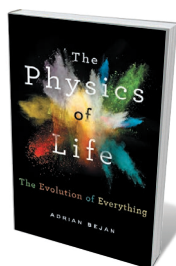
Beekeeping is modish. But as melittologist Thomas Seeley reveals in this captivating study, there is another, uniquely thrilling window on *Apis mellifera*: the sport and science of ‘hunting’ for wild-bee trees. Seeley’s passion for the social insects blazes as he quotes historical accounts by Henry David Thoreau and describes the intricacies of the chase, from baiting with anise-scented sugar syrup to patiently amassing location data. And he delivers the timely reminder that wild honeybee colonies with genetic resilience are key in an era of widespread colony collapse.



Love Canal: A Toxic History from Colonial Times to the Present

Richard S. Newman OXFORD UNIVERSITY PRESS (2016)

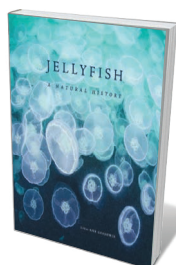
In this chronicle of a notorious US environmental crisis, historian Richard Newman focuses on how community activism forced policy change. In 1978, a health emergency was declared at Love Canal, a suburb of Niagara Falls, New York, when clusters of birth defects and other health issues were discovered — a legacy of chemicals leaching from a 20,000-tonne toxic-waste dump. Protests over the extent and impact of contamination helped to spur the 1980 ‘Superfund’ federal statute that enabled the clean-up of hazardous-waste sites. Newman stresses, however, that the legal and medical saga is not over.



The Physics of Life: The Evolution of Everything

Adrian Bejan ST MARTIN’S (2016)

Energy scientist Adrian Bejan defines life as freely evolving movement in both the inanimate and animate worlds — from a lightning strike to a sprouting seed. This organizational phenomenon is, he argues, underpinned by a principle of physics: “constructal law”, which holds that “power and dissipation conspire to facilitate all movement on earth, animate and inanimate, animal, human, and machine”. Bejan’s treatise crackles with ideas, but seeing analogous patterns in river systems, the spread of ideas and the shift to sustainability can seem a stretch in places.



Jellyfish: A Natural History

Lisa-Ann Gershwin UNIVERSITY OF CHICAGO PRESS (2016)

One resembles an exquisitely ruffled and pleated confection of pale silk chiffon; another, a tangle of bioluminescent necklaces cascading from a bauble. Both marine drifters (*Desmonema glaciale* and *Physalia*) feature in jellyfish expert Lisa-Ann Gershwin’s absorbing coffee-table book on this transparent group with three evolutionary lineages. Succinct science is intercut with surreal portraiture — from the twinkling Santa’s hat jellyfish (*Periphylla periphylla*) to the delicate blue by-the-wind sailor (*Velella velella*). **Barbara Kiser**



MARY GUINAN

Mary Guinan assessed the health of Afghan refugees in camps in Pakistan in the 1980s.

EPIDEMIOLOGY

Chasing epidemics

Tilli Tansey engages with the medical autobiography of a pioneer in the field of HIV/AIDS.

Stories of inspiring female scientists who have cracked the glass ceiling are much in demand. Mary Guinan's *Adventures of a Female Medical Detective* ticks that box — and is a rip-roaring read. An epidemiologist with the US Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia, for decades, Guinan was involved in global smallpox eradication and served as its first female chief scientific adviser. She has also had a distinguished career in HIV/AIDS research as one of the first US scientists to identify early AIDS cases as harbingers of a new epidemic.

As a 'medical detective', Guinan (co-writing with Anne Mather, former managing editor of the CDC newsletter, *Morbidity and Mortality Weekly Report*) presents a series of case studies in explicit homage to super-sleuth Sherlock Holmes. These are not tales of forensic pathology but anecdotes from her career as a field officer for the Epidemic Intelligence Service (EIS), part of the CDC.

Guinan's route to medicine was tortuous. She graduated in chemistry at the turn of the 1960s, when *The New York Times* segregated job adverts by gender, and openings for women in science were few; her first job was developing flavours in a New York City chewing-gum factory. She was rejected by several higher-degree programmes that barred women, finally winning a place at the University of Texas Medical Branch in Galveston,

where she received a PhD in physiology in 1969. Following a secret dream of becoming an astronaut, she enrolled in a space and aviation medicine class at Houston's NASA space centre, only to discover that women were not allowed in the command area, in case they distracted men. Guinan settled on medicine, graduating from Johns Hopkins University in Baltimore, Maryland, in 1972.

Inspired by the World Health Organization's Smallpox Eradication Programme, begun in 1966, she joined the EIS as the only woman in her intake year. But her application for the programme's smallpox work in India was rejected on gender grounds. She insisted that if India could have a female prime minister (Indira Gandhi), the smallpox programme could accommodate a woman. In 1974, she finally succeeded.

A little more reflection on some of Guinan's 'firsts' would be welcome. She was, for instance, sent to investigate bacterial contamination in an intensive-care unit at a unnamed military base. When the commander, announcing that the CDC had sent an expert, asked, "Please will he stand up?", she sat silently. Did she view this as

an effective strategy? How did often being the sole woman doing the work affect her? She mentions a husband and son, but does not discuss whether her domestic situation influenced her career. Some commentary by this successful pioneer would be instructive.

Guinan delivers gripping accounts of work with vulnerable populations. On her first overseas trip for the EIS smallpox programme in 1975, to Uttar Pradesh in India, she lived in rat-infested mud huts while seeking out people with smallpox and their contacts. When the region was declared smallpox-free that May, she decided to dedicate her career to public health.

In 1980, she was part of a CDC group asked by the US state department to assess the health of Afghan refugees in Pakistan. The Soviet Union had invaded Afghanistan, and hostages from the US embassy in Tehran were still in captivity. As Guinan reveals with fury, this risk-ridden situation had a disturbing extra dimension. The CIA, she later learned, was using the CDC team as a front for its operatives. The US government would use this tactic again in 2011, when it created a sham vaccination team to cover its surveillance of al-Qaeda founder Osama bin Laden's house in Pakistan. The repercussions against vaccination programmes and individuals have been catastrophic. As Guinan puts it, the CIA "had stolen part of our souls".

Guinan worked with some of the very first US people with AIDS, and was an expert witness in a landmark legal case that outlawed employment discrimination on the basis of HIV status. Some of her work was captured in Randy Shilt's best-selling book *And the Band Played On* (St Martin's, 1987), later made into a film that disconcerted Guinan with its anodyne portrayal of her. One poignant case study from the mid-1980s is that of 'Lir', a woman infected with HIV by her husband, a preacher jailed for sexually abusing their children. There was no known treatment, and hysteria and condemnation were rife. To keep her HIV status secret in her community, Lir regularly drove more than 190 kilometres to consult Guinan at a clinic in Georgia for several years. In 1995, when Guinan's clinic obtained a supply of the new, effective antiviral drugs, staff members tried to contact all patients likely to benefit. Despite her best efforts, Guinan never found Lir.

With its emphasis on smallpox and AIDS, *Adventures of a Female Medical Detective* will seem to many to fall within the category of medical history. But with Ebola and Zika now in daily headlines, epidemiology and spirited individuals such as Guinan have a very current value for health research. ■

Tilli Tansey is professor of the history of modern medical sciences at Queen Mary University of London.
e-mail: t.tansey@qmul.ac.uk

Adventures of a Female Medical Detective: In Pursuit of Smallpox and AIDS

MARY GUINAN WITH ANNE D. MATHER
Johns Hopkins University Press: 2016.

Correspondence

Spend more on soil clean-up in China

Toxic chemicals from a contaminated site may be a factor in last month's serious sickness among 500 or so students in Changzhou in eastern China. To avoid adverse environmental effects on human health, the country must invest more in soil remediation and create tailored guidelines for decontamination.

Hundreds of thousands of factories have been demolished in China to make way for homes, schools and shopping centres. A national soil survey in 2014 revealed that more than 30% of old industrial land was still polluted (see go.nature.com/a6s5y3; in Chinese). Despite this, the country's total budget for urban soil remediation in 2015 was paltry — roughly equivalent to US\$300 million, or just 0.003% of total gross domestic product.

China's current remediation guidelines for urban soil are based on those of the United States (see go.nature.com/lajaaw; in Chinese). However, the US guidelines were developed mainly for contaminated sites that had been built on, so the same standards may not apply to Chinese brownfield redevelopments. For example, China is more likely to overspend on remediating sites that could have high commercial value.

Such selective remediation, combined with the small budget, could limit the decontamination of urban soils and make it unsustainable.

Yijun Yao Zhejiang University, Hangzhou, China.
yijun_yao@zju.edu.cn

Use open data to curb Zika virus

To avoid losing valuable knowledge and to accelerate decision-making during the current Zika public-health emergency, the World Health Organization (WHO) and international partners are

renewing efforts to promote rapid sharing of the latest research data (see go.nature.com/qtf5x4). Data sharing is important for all medical research, and particularly during outbreaks of such new and unstudied diseases (see, for example, N. L. Yozwiak *et al.* *Nature* **518**, 477–479; 2015).

The International Committee of Medical Journal Editors has ruled that, in a WHO public-health emergency, dissemination of raw information critical to public health will not prejudice later publication by researchers in the same journal. This is important, because it prioritizes open access and real-time disclosure over competition between researchers and companies rushing to publish successful trial results.

The *Bulletin of the World Health Organization* has also made publication of papers and raw data on Zika virus open-access and immediate (see go.nature.com/djfyzf). So far, more than 30 funding and research agencies and medical journals are supporting the initiative.

Failure to disclose medical research data promptly and publicly can give rise to misinformation, leading to treatments that are dangerous or ineffective, or to delays in effective treatments. It also wastes precious public-health resources. **Marie-Paule Kieny, Vasee Moorthy, Daniela Bagozzi** WHO, Geneva, Switzerland.
bagozzid@who.int

Restoration: avoid arbitrary baselines

Janne Kotiaho and colleagues propose using a pre-degradation 'natural state' as a reference baseline for assessing the impact of humans on biodiversity and ecosystem function (*Nature* **532**, 37; 2016). However, it is not possible for scientists to define a single such baseline objectively. This is because global ecosystems changed drastically during pre-human time periods, under

otherwise 'natural' conditions.

As long as scientists set baselines to single time points in Earth's history, this problem will remain. The established solution is to compare the full ranges of variability in biodiversity and ecosystem functioning within and between pre-human and post-human worlds (see K. J. Willis and H. J. Birks *Science* **314**, 1261–1265; 2006).

This approach avoids the need to set any arbitrary baselines for a 'natural state'. It also allows scientists to determine what effects human activities (such as nitrogen pollution or greenhouse-gas emissions) have on the planet, after expected ranges of natural changes have already been accounted for (see K. K. McLauchlan *et al.* *Nature* **495**, 352–355; 2013).

Zia Mehrabi University of British Columbia, Canada.
zia.mehrabi@ubc.ca

Restoration: 'Garden of Eden' unrealistic

We consider the proposed use of a 'pre-degradation' state as a reference baseline for damaged ecosystems to be unrealistic (J. Kotiaho *et al.* *Nature* **532**, 37; 2016). Instead of this 'Garden of Eden' baseline, we argue that restoration should respond to current drivers of biodiversity loss and decline in ecosystem function and services.

A baseline that prescribes a list of pre-degradation species is a good place to start, but it does not take into account the dynamism of ecological communities, in which species are constantly migrating, evolving and going extinct. Moreover, native species can be difficult to propagate and invasive species may be so prevalent that they are impossibly costly to remove. Present-day climate change may necessitate the use of non-local genotypes and even non-local native species to improve restoration outcomes (see M. F. Breed *et al.* *Conserv. Genet.* **14**, 1–10; 2013

and R. J. Hobbs *Rest. Ecol.* **24**, 153–158; 2016).

We suggest that restoration efforts should focus on a trajectory towards functional, self-sustaining ecosystems that are resilient to climate change and provide measurable ecosystem-service outcomes — as emphasized by the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES). **Martin F. Breed, Andrew J. Lowe** University of Adelaide, Australia.
Peter E. Mortimer Kunming Institute of Botany; and World Agroforestry Centre, Kunming, China.

martin.breed@adelaide.edu.au

Shark-fin landing policy aids control

David Sims and Nuno Queiroz call for tighter fisheries regulations for species caught by European fleets as by-catch, using shortfin mako and blue sharks as examples (*Nature* **531**, 448; 2016). However, their arguments with respect to these species have been overtaken by policy developments.

In 2013, the European Parliament and the Council of the European Union adopted a new regulation that amends a 2003 legal act about the removal of fins of sharks on board vessels. Sharks must now be landed with their fins attached, so gutted carcasses of swordfish can no longer be passed off as shortfin mako.

As an indicator of that policy's success, actual shortfin mako landings of the EU fleet made up 16.5% and 9.7% of blue-shark landings in 2013 and 2014, respectively (go.nature.com/6fptm8), in line with the typical proportions quoted by Sims and Queiroz.

Alexander J. Stein European Commission, Directorate-General for Maritime Affairs and Fisheries, Brussels, Belgium.
alexander.stein@ec.europa.eu
Disclaimer declared (see go.nature.com/yriqgt for details).

Harry Kroto

(1939–2016)

Discoverer of new forms of carbon.

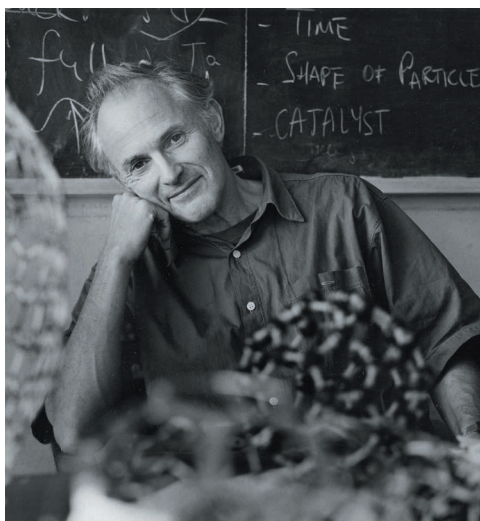
Harry Kroto was part of the team that discovered buckminsterfullerene, the football-shaped carbon-60 molecule that came to be known as a buckyball. The realization that such a large molecule could self-assemble from hot carbon vapour forced a reassessment of the science of carbon. By prompting searches for other structures — carbon nanotubes and nanowires were among the materials later found — the discovery ultimately provided a foundation for nanoscience and nanotechnology.

Kroto, who died on 30 April, was born Harold Krotoschiner in 1939 in Wisbech, UK, the son of German refugees. During the Second World War, his father Heinz was interned on the Isle of Man as an enemy alien, and Kroto and his mother Edith went to live in the town of Bolton. After the war, the family stayed in Bolton, where his father opened a balloon factory and shortened the family name to Kroto. After finishing school, Kroto studied chemistry at the University of Sheffield, completing his undergraduate degree in 1961 and his PhD in 1964.

After two postdoc positions, Kroto and his wife Margaret moved to Brighton in 1967, where he took up a teaching post at the University of Sussex. During the 1970s, Kroto began work that led to a lifelong fascination with the chemistry of interstellar space.

Kroto studied a class of linear molecules called cyanoacetylenes with the atomic composition $\text{H}-(\text{C})_n-\text{C}\equiv\text{N}$. These were hypothesized to exist in the molecular clouds that surround stars with carbon-rich atmospheres. He combined molecular spectroscopy of these carbon chains — which he created fleetingly in the lab — with measurements made by radioastronomers who were aiming to detect the same molecules in the material surrounding carbon stars. By the early 1980s, radioastronomers had detected cyanoacetylenes in space that contained up to 11 carbon atoms. The relatively high abundance of these molecules challenged existing models of interstellar chemistry, which predicted the presence of much smaller molecules.

Around this time Kroto learned about Richard Smalley's cluster-beam experiments. Smalley, then a physical chemist at Rice University in Houston, Texas, was using a laser to vaporize a material with a high melting point (called a refractory target). As the hot atoms cooled, they would condense. Smalley halted this chemistry after only a few microseconds using supersonic



expansion: a high-pressure gas was passed through a small orifice into a large vacuum chamber to cool the molecules and stop all chemical reactions.

The technique offered the perfect way to test whether vaporized carbon would condense to form carbon chains similar to those found in certain interstellar environments. So in 1985, Kroto travelled to Houston to work in Smalley's lab. It was during this visit that he, along with Smalley and his group (including the two of us), discovered C_{60} and the other fullerenes. And Kroto was able to prove his hypothesis that long carbon chains were reaction products of condensing carbon.

Before that, three different crystalline forms, or allotropes, of carbon were known: graphite, diamond and lonsdaleite, a rare modification of diamond. The first two provided text-book examples of how physical properties reflect atomic structure: the electron arrangement in graphite (sp^2 -hybridized) makes the allotrope an electrical conductor and a dry lubricant; whereas that of diamond (sp^3 -hybridized) makes it an insulator and the hardest known mineral. Equally fundamental, the molecular chemistry of carbon provides the foundation of organic chemistry and biochemistry.

It is thus not surprising that our proposed structure for C_{60} — a truncated icosahedron in which the 60 carbon atoms form a cage of interlocking pentagons and hexagons — was initially viewed with scepticism (H. W. Kroto *et al. Nature* **318**, 162–163; 1985). Our experimental support for the C_{60} structure

arose from a combination of mass-spectra data and circumstantial evidence. This was hardly the gold standard of single-crystal X-ray analysis for absolute molecular structure determination. However, the football structure followed Occam's razor: it tied together many observations in a simple and elegant way, and yielded many predictions that were later proved to be correct, including the structure of a second fullerene, C_{70} .

Absolute confirmation of these structures came five years later, when physicists Don Huffman and Wolfgang Krätschmer and their groups worked out how to make C_{60} in bulk. Today, the buckyball is a crucial component of solar cells.

In 1996, Kroto shared the Nobel Prize in Chemistry with Smalley and one of us (R.F.C.), and was knighted. From 2002 to 2004, he served as president of the Royal Society of Chemistry, and in 2004 he left Sussex to take up a chair at Florida State University in Tallahassee.

Harry was strongly opinionated. He did not profess modesty, and as an atheist, he would often engage his religious acquaintances in fierce debate. But with children (he had two sons), he was always terrific. After receiving the Nobel, he devoted much of his time to elevating the importance of science teaching. Seeing that the football-like structure of C_{60} would resonate with almost any child, he would set up games in which buckyballs would pop out of unexpected places, or have children assemble buckyballs themselves.

Harry had an impish sense of humour similar to that of the British comedy group Monty Python, which he greatly admired. He also had the distinction of being the only Nobel laureate to have appeared on stage with the actor Ian McKellen — in a school production when they were both teenagers.

Harry (who remained friends with McKellen) had a deep appreciation of the arts, and was himself a skilled graphic artist. He published several designs, one of which was chosen for the 2001 UK postage stamp celebrating the Nobel centenary. Of course, it included a drawing of a buckyball. ■

James R. Heath is professor of chemistry at the California Institute of Technology, Pasadena, California, USA. **Robert F. Curl** is emeritus professor of chemistry at Rice University, Houston, Texas, USA. e-mails: heath@caltech.edu; rfcurl@rice.edu

ANNE PURWISS/THE ROYAL SOCIETY

CELL BIOLOGY

Choreography of protein synthesis

Both nuclear genes and genes in organelles called mitochondria are involved in the assembly of the cellular energy-producing machinery. RNA-translation programs that coordinate the two systems have now been identified. [SEE ARTICLE P.499](#)

MARTIN OTT

In cells, organelles known as mitochondria convert chemical energy from food into ATP, a molecule that fuels most of the reactions of life. Energy conversion relies on a series of macromolecular machines collectively called the oxidative phosphorylation system, which consists of protein complexes of the respiratory chain and the large enzyme ATP synthase. These complexes are composed of protein subunits that are encoded by either nuclear or mitochondrial DNA. In this issue, Couvillion *et al.*¹ (page 499) provide insight into how these two genetic systems are coordinated in time, despite being separated physically by cellular membranes.

The two distinct genetic systems in eukaryotic cells (cells that are characterized by membrane-bounded compartments) have their origin in an endosymbiotic event that occurred roughly 1.5 billion years ago — when a bacterial cell and an archaeal cell merged. This resulted in combining the efficient energy-conversion system of mitochondria (the former bacterial cells) with the greater complexity of the original archaeal cell. During evolution, most of the former bacterial genes were transferred to the nuclear genome². This in turn created the need for systems to import nuclear-encoded proteins into mitochondria³.

Mitochondria still contain a vestigial genome that encodes a limited set of proteins, including key subunits of the oxidative phosphorylation system. These proteins are synthesized by mitochondrion-specific ribosomes (RNA-protein complexes that mediate protein synthesis), which have developed from the endosymbiont's bacterial ribosome into a particle that differs in structure and composition from the ribosomes in the cytoplasm^{4,5}. The dual genetic origin of subunits of the oxidative phosphorylation complexes necessitates tight coordination between mitochondrial- and nuclear-gene expression to supply similar quantities of subunits for the assembly process (Fig. 1).

Couvillion and colleagues performed their analyses in yeast. When yeast cells switch from obtaining their energy through the anaerobic fermentation of glucose to oxygen-requiring respiration, dramatic reprogramming of

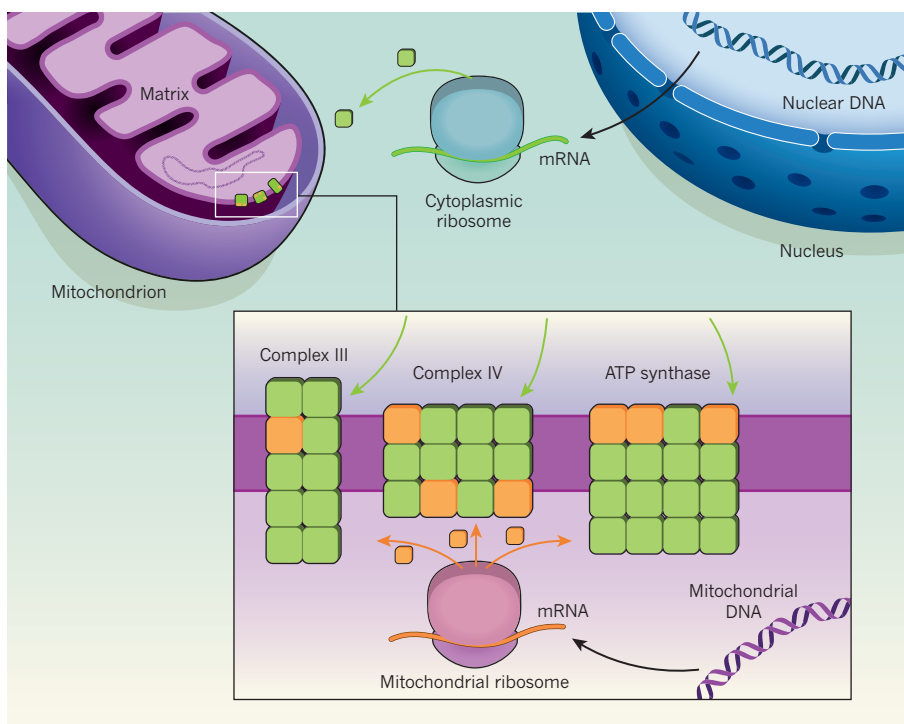


Figure 1 | Composite protein assembly in mitochondria. The mitochondrial oxidative phosphorylation system consists of a series of large complexes (such as complexes III and IV and ATP synthase), each containing many subunits. The subunits are encoded by mitochondrial or nuclear genes. Messenger RNAs that encode subunits with a nuclear origin are translated by ribosomes in the cytoplasm, and the resulting proteins are imported into mitochondria. By contrast, mRNAs of the subunits encoded by mitochondrial genes are translated by mitochondrial ribosomes in the organelle's matrix. Couvillion *et al.*¹ report that the translation, but not the transcription, of nuclear and mitochondrial mRNAs is synchronized.

gene expression occurs. This phenomenon has been studied extensively in the context of alterations in the expression of nuclear genes. Couvillion and colleagues' study, however, sets a new standard by including analyses of mitochondrial-gene expression, revealing previously unrecognized complexities and levels of regulation.

The classical view of gene reprogramming is that changes in metabolism and its accompanying stress activate the transcription of specific genes and so provide increased amounts of the associated messenger RNAs. Couvillion *et al.* combined analyses of transcription with quantification of the efficiency with which these mRNAs are translated into proteins, using an approach called ribosome profiling. Specifically, they performed

profiling on the cytoplasmic and mitochondrial translation machineries in parallel.

The authors found that the metabolic shift from fermentation to respiration in yeast resulted in a rapid accumulation of all of the nuclear-transcribed mRNAs that encode subunits of the oxidative phosphorylation system. Surprisingly, however, the translation of these mRNAs did not increase equally for all transcripts. Transcripts that encode subunits constituting respiratory-chain complexes (such as complexes III and IV) were preferentially translated, whereas translation of those encoding ATP-synthase subunits was repressed.

Another unexpected finding was that mitochondrial protein synthesis followed the same translational program as its cytoplasmic counterpart, with subunits of the respiratory

chain gaining translational efficiency at the expense of ATP synthase subunits. In essence, therefore, the two genetic systems respond identically, despite being located in different compartments.

The authors also found that changes in the translation of nuclear-encoded mRNAs in the cytoplasm were independent of mitochondrial translation. By contrast, inhibition of cytoplasmic ribosomes not only induced the translation of many mitochondrial transcripts, but also specifically reduced the synthesis of some proteins. The latter observation is in line with previously identified feedback mechanisms^{6–8} that adjust the synthesis of a subset of mitochondrial proteins to levels that can be assembled into oxidative phosphorylation complexes.

Although the reprogramming of cytoplasmic translation during stress is well documented⁹,

how mitochondrial translation is adjusted in response to stress and altered metabolic needs is largely unknown. The expression of individual mitochondrial mRNAs is controlled by translational activators¹⁰ — a diverse family of nuclear-encoded RNA-binding proteins with ill-defined molecular functions. Future challenges therefore include unravelling the exact molecular functions of these translational activators and how they cooperate with other factors involved in translation initiation, to explain how metabolic cues modulate protein synthesis in mitochondria. An equally exciting challenge will be to extend this research from yeast to more-complex eukaryotic cells, such as those of mammals. ■

Martin Ott is at the Center for Biomembrane Research, Department of Biochemistry and

Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden.

e-mail: martin.ott@dbb.su.se

1. Couvillion, M. T., Soto, I. C., Shipkovenska, G. & Churchman, L. S. *Nature* **533**, 499–503 (2016).
2. Archibald, J. M. *Curr. Biol.* **25**, R911–R921 (2015).
3. Neupert, W. J. *Mol. Biol.* **427**, 1135–1158 (2015).
4. Greber, B. J. & Ban, N. *Annu. Rev. Biochem.* <http://dx.doi.org/10.1146/annurev-biochem-060815-014343> (2016).
5. Ott, M., Amunts, A. & Brown, A. *Annu. Rev. Biochem.* <http://dx.doi.org/10.1146/annurev-biochem-060815-014334> (2016).
6. Barrientos, A., Zambrano, A. & Tzagoloff, A. *EMBO J.* **23**, 3472–3482 (2004).
7. Rak, M. & Tzagoloff, A. *Proc. Natl Acad. Sci. USA* **106**, 18509–18514 (2009).
8. Gruschke, S. et al. *J. Cell Biol.* **199**, 137–150 (2012).
9. Sonenberg, N. & Hinnebusch, A. G. *Cell* **136**, 731–745 (2009).
10. Costanzo, M. C. & Fox T. D. *Annu. Rev. Genet.* **24**, 91–113 (1990).

This article was published online on 11 May 2016.

ASTROPHYSICS

How black holes restrain old galaxies

Supermassive black holes are thought to keep star formation under control by ejecting or stirring gas in galaxies. Observations of an old galaxy reveal a potential mechanism for how this process occurs. SEE LETTER P.504

MARC SARZI

When supermassive black holes at the centres of galaxies accrete matter, they turn into powerful engines that can potentially expel the gas of their host galaxies, thereby halting star formation¹. Despite numerous efforts^{2,3} to observe supermassive black holes in the process of quenching star formation, conclusive evidence for such a process has remained elusive, particularly in the nearby Universe. On page 504 of this issue, Cheung *et al.*⁴ present state-of-the-art observations that might finally show how supermassive black holes can prevent galaxies that are already dominated by old, red stars from forming new ones.

According to our current view of their formation, galaxies grow by merging with other galaxies or by forming new stars, either from freshly acquired gas or from material that is lost by their old and dying stars. Merging events can further rearrange them into rounder shapes, whereas galaxies that benefit from a constant supply of external or recycled gas can form a stellar disk dominated by young, blue stars. Galaxies that have no external gas supply evolve passively into old, red stellar systems collectively called early-type galaxies (otherwise known as elliptical or lenticular galaxies).

And yet, up to 75% of early-type galaxies

contain gas that could potentially fuel new bursts of star formation⁵. The fact that stars are observed to form in only 10–20% of such galaxies^{6,7} suggests that some kind of star-formation quenching is taking place. This is

consistent with the finding that gas in early-type galaxies is usually in a warm, ionized state, rather than existing as cold clouds of gaseous molecules from which stars can form.

The radiation emitted from ionized gas is generally powered by hot but old stars^{8,9}, rather than by massive, newly born stars such as those in the disk of the Milky Way and in other spiral galaxies. Also, unlike cold molecular gas, which always orbits in a thin disk at the circular velocity set by the local gravitational potential, the warm gas of early-type galaxies often shows sizeable, random motion — suggesting either that it is being stirred up somehow, or has yet to settle down¹⁰. However, the kinematics of the ionized gas in early-type galaxies has so far been considered to be consistent with coherent, although perhaps not completely ordered, rotation.

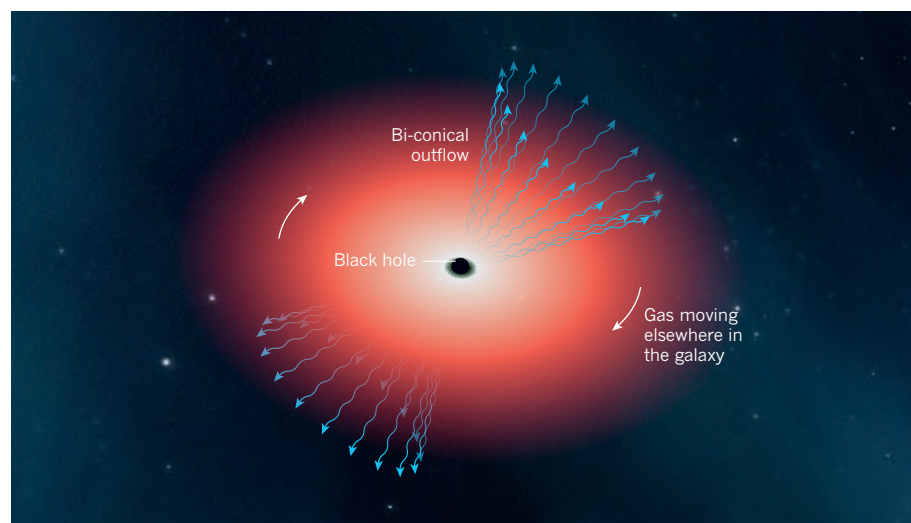


Figure 1 | Bi-conical gas outflow from an early-type galaxy. Cheung *et al.*⁴ propose a model of gas outflow from an old, red galaxy (an early-type galaxy) that explains their observations of an early-type galaxy nicknamed Akira. The authors suggest that, when the supermassive black hole at the centre of such a galaxy accretes matter, this can stimulate nuclear activity that drives a bi-conical outflow of gas — stirring up gas that is moving elsewhere in the galaxy, and possibly expelling part of it. This provides a possible mechanism by which star formation in early-type galaxies is quenched.

Enter Cheung and colleagues. The authors used spectroscopic observations that allowed them to map the motion of ionized gas across a galaxy and to infer what is powering the gas's emission. In this way, they conclusively show that ionized gas in a substantial fraction of early-type galaxies does not rotate in a coherent fashion. Instead, the authors propose a model whereby the approaching and receding material that is observed across the gas-velocity field of such objects is due to a bi-conical outflow of gas powered by a centrally active supermassive black hole (Fig. 1), rather than resulting from the circulation of gas in an inclined disk.

In maps of the intensity of ionized-gas emission, the class of early-type galaxy in which such bi-conical outflows occur is characterized by bisymmetrical, elongated features that align with the gradient in the gas-velocity field. Cheung and colleagues' model has the advantage of having a simple explanation for this defining characteristic: it reflects the accumulation of material on both the approaching and receding sides of the outflow.

The researchers present observations for one prototypical example of this class of object, in which the activity of the central supermassive black hole seems to have been triggered by interaction with a nearby companion galaxy. In this example, which Cheung *et al.* nickname Akira, such activity is sufficient to sustain the kinetic power of the outflowing wind, which in turn balances the cooling of the warm, ionized gas. Even if the central activity is not enough to rid Akira of its gas, it would provide sufficient energy to stir it by causing turbulence and shocks, and therefore still prevent the gas cooling that leads to star formation.

Although the bi-conical-outflow model presented by Cheung and co-workers is only qualitative, the authors' results might aid our understanding of the role of supermassive black holes in galaxy evolution. Akira is just one of the 10,000 objects that will eventually be targeted by the ongoing campaign (the MaNGA survey¹¹) from which the authors' data are drawn. From the 700 MaNGA galaxies presently being surveyed, Akira-like objects occur as a small (5%), yet non-negligible fraction of early-type galaxies⁴. This could be just the tip of the iceberg — central-black-hole activity can be triggered several times by different accretion episodes, and thus many galaxies that do not presently show outflows could have been stirred, or their gas expelled, by a previous episode.

The observed outflows may also help to solve another riddle: the origin of gas in early-type galaxies. One way to tell whether an early-type galaxy acquired gas from other galaxies or from recycled material lost from its stars is to compare the gas's angular momentum with that of the stars. If the gas was internally produced, it should follow the motions of the stars, whereas if it was externally acquired

it could just as well move in the opposite direction. Observations¹² of the stellar and gaseous kinematics of early-type galaxies show that the gas comes from mixed sources. This is puzzling, but not really problematic, given that, for instance, galaxies in crowded environments such as galaxy clusters do not interact easily with each other and therefore find it hard to steal gas from smaller companions¹².

More troubling is the fact that 25% of early-type galaxies have little or no gas at all⁵. Because early-type galaxies have similar, old stellar populations that would also return gas to their hosts over time, one would expect all early-type galaxies to retain at least some of this recycled material in the absence of an external gas source. By providing evidence for a mechanism capable of removing at least part of the gas, Cheung and colleagues' work might bring us a step closer to explaining why some early-type galaxies seem to be devoid of gas. ■

CELL BIOLOGY

Killer enzymes tethered

Caspase enzymes promote cell death, but are also involved in sperm development in fruit flies. The discovery that, in sperm, caspase activation is restricted to the surface of organelles called mitochondria sheds light on this unusual role.

SHIGEKAZU NAGATA

Protease enzymes called caspases are renowned killers, cleaving proteins to execute a program of apoptotic cell death. As such, the discovery¹ in 2003 that caspase activation is required for sperm differentiation in the fruit fly *Drosophila melanogaster* came as a surprise. Writing in *Developmental Cell*, Aram *et al.*² report that the restriction of caspase activity to the surfaces of organelles called mitochondria allows the enzymes to exert this alternative effect.

During the development of *Drosophila* sperm, precursors called spermatids that are linked to one another by cytoplasmic bridges mature simultaneously. Their nuclei elongate, their mitochondria fuse to form two large aggregates, new organelles form and membranes are added around each sperm cell³. At the end of this process, organelles and cytoplasmic materials that are not needed in the mature sperm are removed in vesicles, and the spermatids separate from each other in a process called individualization.

Individualization and disposal of cytoplasmic material both proceed from head to tail along spermatids, and caspases are activated in a head–tail gradient. The discovery that caspase inhibitors block individualization and cause male sterility provided the first evidence that caspases

Marc Sarzi is at the Centre for Astrophysics Research, University of Hertfordshire, Hatfield AL10 9AB, UK.

e-mail: m.sarzi@herts.ac.uk

1. Fabian, A. C. *Annu. Rev. Astron. Astrophys.* **50**, 455–489 (2012).
2. Schawinski, K. *et al. Mon. Not. R. Astron. Soc.* **382**, 1415–1431 (2007).
3. Ciccone, C. *et al. Astron. Astrophys.* **562**, A21 (2014).
4. Cheung, E. *et al. Nature* **533**, 504–508 (2016).
5. Sarzi, M. *et al. Mon. Not. R. Astron. Soc.* **366**, 1151–1200 (2006).
6. Sarzi, M. *et al. Mon. Not. R. Astron. Soc.* **402**, 2187–2210 (2010).
7. Young, L. M. *et al. Mon. Not. R. Astron. Soc.* **414**, 940–967 (2011).
8. Binette, L., Magris, C. G., Stasińska, G. & Bruzual, A. G. *Astron. Astrophys.* **292**, 13–19 (1994).
9. Stasińska, G. *et al. Mon. Not. R. Astron. Soc.* **391**, L29–L33 (2008).
10. Young, L. M., Bureau, M. & Cappellari, M. *Astrophys. J.* **676**, 317–334 (2008).
11. Bundy, K. *et al. Astrophys. J.* **798**, 7 (2015).
12. Davis, T. A. *et al. Mon. Not. R. Astron. Soc.* **417**, 882–899 (2011).

were involved in sperm development¹.

In 2007, the biologist Eli Arama and colleagues screened sterile male flies for mutants that could not activate caspases during sperm individualization⁴. This revealed that a testis-specific enzyme called Cullin-3-based ubiquitin ligase (CRL3), which attaches ubiquitin molecules to proteins to modify the proteins' behaviour, is required for caspase activation. The group proposed that CRL3 ubiquitinates an apoptosis-inhibiting protein called Bruce, which is then degraded, enabling caspase activation.

Subsequently, Arama's laboratory identified a protein called Soti that functions as a CRL3 inhibitor⁵ by competing with CRL3 targets to bind to the enzyme. Soti is concentrated in the tail region of spermatids and its levels form a gradient opposite to that of activated caspases. Thus, it has been proposed that CRL3 determines the level of activated caspase and that Soti inhibits caspase activation⁵. This mechanism explains how a caspase gradient forms during individualization, but why activated caspase does not kill spermatids has remained a mystery.

In the group's latest study, Aram *et al.* found that a testis-specific version of the mitochondrial enzyme succinyl-CoA synthetase (SCS) mediates caspase activity in spermatids. SCS is a key enzyme in a process called the Krebs cycle, which generates energy in all

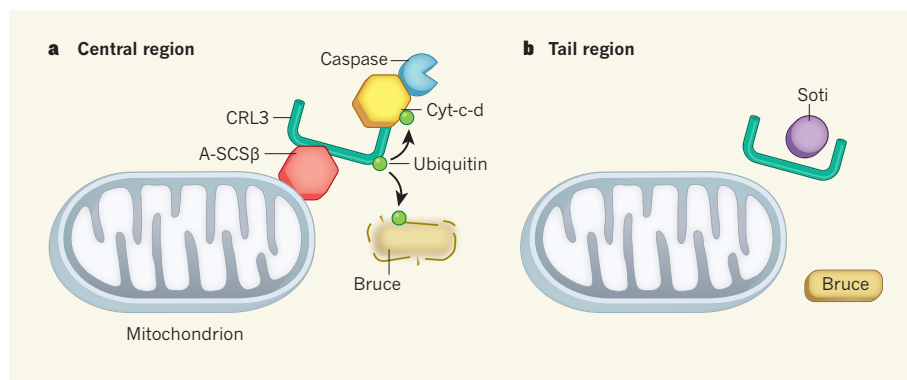


Figure 1 | Caspase-mediated sperm development. Protease enzymes called caspases normally trigger apoptotic cell death, but low-level activity promotes sperm development in fruit flies. **a**, Aram *et al.*² report that, in the central region of sperm, the Krebs-cycle enzyme A-SCSβ acts as an anchor that tethers the CRL3 ubiquitin ligase enzyme to the surface of mitochondria, and as an inhibitor of the CRL3-inhibiting protein Soti (not shown). CRL3 adds ubiquitin molecules to the caspase-activating protein cytochrome *c* (Cyt-*c-d*) and to the apoptosis-inhibiting protein Bruce. The authors speculate that this ubiquitination leads to degradation of Bruce, and activation of Cyt-*c-d* only in the region around the mitochondria enables it to activate caspases in a localized area without causing apoptosis. **b**, In the tail of sperm, Soti binds to CRL3, preventing caspase activation, and Bruce is not degraded.

aerobic organisms. The enzyme has α and β subunits, and the β subunit exists in two forms, which determine whether the enzyme synthesizes energy-carrying ATP (A-SCSβ) or GTP molecules (G-SCSβ) from the Krebs-cycle intermediate succinyl-CoA.

Aram and colleagues demonstrated that testis-specific A-SCSβ has a Krebs-cycle-independent role in caspase activation. They found that levels of A-SCSβ were elevated at the onset of spermatid individualization, and that the protein serves as an anchor to tie CRL3 to the surface of mitochondria (Fig. 1). Moreover, A-SCSβ prevented Soti from binding to CRL3, blocking its inhibitory activity. It seems, therefore, that A-SCSβ keeps caspase activity at low levels and limits it to a restricted zone in the central region of the cylinder-like spermatids. This prevents the proteins from spreading to the nucleus and plasma membrane, where many caspase targets are located.

This intriguing proposal is puzzling for several reasons. First, given that A-SCSβ is normally located within mitochondria, what mediates its release to the external surface? In mammalian apoptosis, proteins of the Bcl-2 family stimulate the release from mitochondria of another apoptosis-promoting protein, cytochrome *c*, which can then activate caspases. Although Bcl-2 family members do not seem to play a major part in fruit-fly apoptosis, some members are required for the death of around one-third of early-stage sperm precursors called spermatogonia as a normal part of sperm maturation⁶. Thus, Bcl-2 family proteins might also be involved in releasing A-SCSβ from spermatid mitochondria.

Another Krebs-cycle enzyme, fumarase, also has both mitochondrial and cytoplasmic functions⁷. Like A-SCSβ, fumarase is synthesized in the cytoplasm, and transported into

mitochondria thanks to a mitochondrion-targeting sequence in its amino-terminal domain. However, some fumarase molecules are thought to return to the cytoplasm during the import process⁸. Similarly, some of the A-SCSβ that is synthesized at the onset of sperm individualization might remain at the mitochondrial surface.

Another question is how CRL3 activates caspases. As previously proposed⁴, CRL3-mediated ubiquitination of Bruce, followed by Bruce degradation, could be the mechanism for caspase activation. Another CRL3 target is cytochrome *c* (ref. 2). In mammals, cytochrome *c* — in complex with a scaffold protein — cleaves procaspase proteins into active caspases. In fruit flies, testis-specific cytochrome *c* (Cyt-*c-d*) activates caspases in spermatids¹. Because ubiquitination can regulate protein–protein interactions⁹, perhaps ubiquitinated, but not de-ubiquitinated, Cyt-*c-d* can activate caspases.

During sperm individualization, cytoplasmic material is disposed of in vesicular waste bags. Cullin-mediated ubiquitination is involved in establishing the structure of an organelle called the Golgi, which packages cargo into vesicles and also has a role in sorting the proteins for vesicle packaging¹⁰. Because several proteins in the Golgi (and in the other organelles involved in vesicle transport, such as the endoplasmic reticulum and lysosomes) are targets of caspases, cytoplasmic protein disposal may be collaboratively controlled by CRL3-mediated ubiquitination and caspase-mediated protein cleavage.

Finally, the non-apoptotic activation of caspases has been observed in many other biological processes, in both vertebrates and invertebrates¹¹. It will be fascinating to discover whether similar mechanisms or molecules are involved in these processes. ■



50 Years Ago

Termites show great activity around even a small breach in their nest and soon begin to build to repair the damage ... I have investigated the nest-building behaviour of the damp-wood termites *Zootermopsis angusticollis* and *Z. nevadensis* ... The evidence obtained ... shows that a very small proportion of termites can detect air movements of the order of one thousandth of those in a closed room. The sense organs concerned are believed to be located on the antennae ... This very high sensitivity to air movement clearly helps to explain how termites are attracted to even minute openings of their nest ... It also suggests great caution is necessary in the design of experiments with termites.

From *Nature* 28 May 1966

100 Years Ago

The word “blizzard,” signifying originally a type of snowstorm most common and most severe in the Rocky Mountain States of the Union, although occasionally occurring elsewhere, is now loosely used to mean any heavy snowstorm. This is unfortunate, for a term is needed for the type of storm referred to above. Three things must co-exist in a blizzard—large quantities of very fine snow; very low temperature, generally below zero Fahrenheit; and a high wind of great velocity.

Apparently the loose use of the word is becoming common in Great Britain, for you refer in *NATURE* of April 6 (p. 129) to “a blizzard of unusual severity.” The context shows that neither the snow nor the temperature condition could have been fulfilled, for you say that the gale “was accompanied by rain and snow.”

I doubt very much whether the British Isles could produce the requisite conditions for a real blizzard.

From *Nature* 25 May 1916

Shigekazu Nagata is in the Laboratory of Biochemistry and Immunology, WPI Immunology Frontier Research Center, Osaka University, Osaka 565-0871, Japan. e-mail: snagata@ifrec.osaka-u.ac.jp

1. Arama, E., Agapite, J. & Steller, H. *Dev. Cell* **4**, 687–697 (2003).
2. Aram, L. *et al.* *Dev. Cell* **37**, 15–33 (2016).
3. Fabian, L. & Brill, J. A. *Spermatogenesis* **2**, 197–212 (2012).
4. Arama, E., Bader, M., Rieckhof, G. E. & Steller, H. *PLoS Biol.* **5**, e251 (2007).
5. Kaplan, Y., Gibbs-Bar, L., Kalifa, Y., Feinstein-Rotkopf, Y. & Arama, E. *Dev. Cell* **19**, 160–173 (2010).
6. Yacobi-Sharon, K., Namdar, Y. & Arama, E. *Dev. Cell* **25**, 29–42 (2013).
7. Monaghan, R. M. & Whitmarsh, A. J. *Trends Biochem. Sci.* **40**, 728–735 (2015).
8. Yogev, O., Naamati, A. & Pines, O. *FEBS J.* **278**, 4230–4242 (2011).
9. Mukhopadhyay, D. & Riezman, H. *Science* **315**, 201–205 (2007).
10. Lu, A. & Pfeffer, S. R. *Trends Cell Biol.* **24**, 389–399 (2014).
11. Kuranaga, E. & Miura, M. *Trends Cell Biol.* **17**, 135–144 (2007).

This article was published online on 18 May 2016.

EVOLUTION

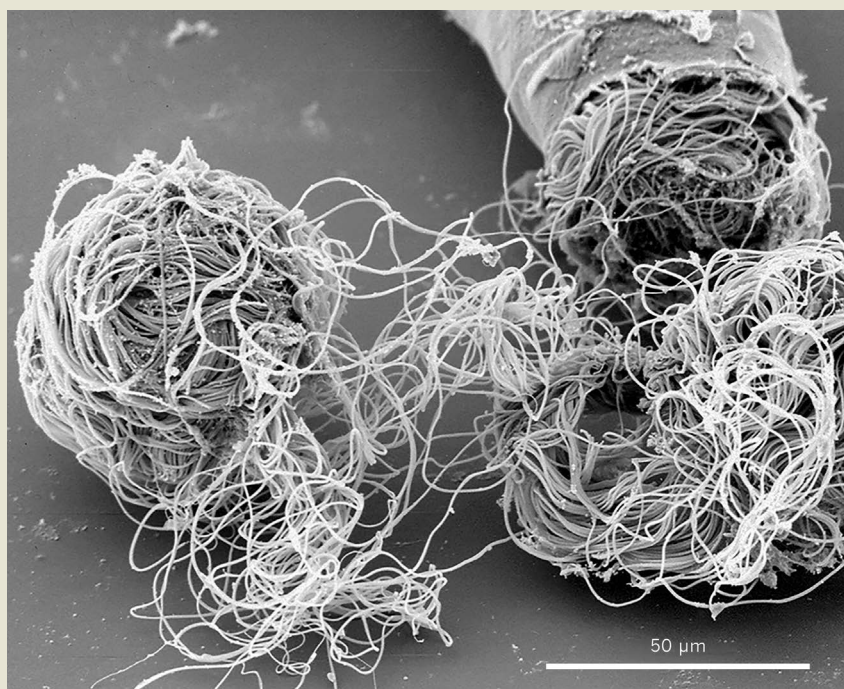
The bigger, the better

The sperm of some species of fruit fly are up to 5.8 centimetres long — around 20 times as long as the fly itself (pictured, two *Drosophila bifurca* sperm spill out of the male's ruptured seminal vesicle). Giant sperm tails are energetically expensive to produce, and their evolution requires intensive sexual selection. And yet theory predicts that sexual selection should steadily weaken as sperm get larger and their numbers decline. On page 535, Lüpold *et al.* investigate this 'big-sperm paradox' (S. Lüpold *et al.* *Nature* **533**, 535–538; 2016).

Longer sperm are better than their shorter comrades at displacing competitors from the female's seminal receptacle, and the length of the seminal receptacle determines the female's preference for

longer sperm. The authors demonstrate that the genes that confer longer sperm are correlated with those that confer a longer seminal receptacle. Thus, female preference and sperm size coevolve. This process is reinforced by the fact that a longer seminal receptacle correlates with shorter intervals between female mating, increasing sperm competition and hence the advantage to longer sperm.

But what is the benefit to females? Only large, healthy males that carry 'good' genes can produce long sperm in sufficient quantities to outcompete other suitors. Therefore, a long seminal receptacle maximizes a female's chances of producing offspring that have high-quality genes. Both sexes win in this evolutionary arms race. **Jennifer R. Gardiner**



ROMANO DALLAI

DEPRESSION

Ketamine steps out of the darkness

The way in which ketamine exerts its antidepressant effects has been perplexing. Evidence that a metabolite of the drug is responsible, and acts on a different target from ketamine, might be the key to an answer. [SEE ARTICLE P.481](#)

ROBERTO MALINOW

The novelist William Styron, who experienced depression, referred to the disorder as a black and howling tempest in the brain, noting¹ that “the wisest books among them underscore the hard truth that serious depressions do not disappear overnight”. Indeed, depression is a painful and often deadly disorder that frequently requires months or more of treatment and that, for around one-third of sufferers, is treatment-resistant². Ketamine is an attractive therapeutic, because it can act rapidly and effectively against even treatment-resistant depression^{3–6} — but the drug has side effects and does not always work. An understanding of ketamine’s mechanism of action, which could lead to improved treatments, has been widely sought. In this issue, Zanos *et al.*⁷ (page 481) provide several lines of evidence to indicate that it is not ketamine itself, but one of its metabolites, that is responsible for the drug’s antidepressant effects.

Ketamine has a moderately high binding affinity for, and can block the activity of, the NMDA receptor protein (NMDAR)⁸. This receptor is perhaps best known for its requirement⁹ in a phenomenon called long-term potentiation (LTP), which occurs widely in the brain, whereby the synaptic connections between neurons are strengthened, enhancing neural signalling¹⁰. The enhanced signalling produced by LTP underlies the formation of associative memories^{11,12}.

How can transient blockade of NMDAR, and possibly LTP, have a rapid and long-lasting effect on human depression? Given the role of LTP in memory formation, it might be logical to assume that ketamine causes a brief block in the formation of memories. But even if this were true, how could it alleviate depression? To many physiologists, the idea that blocking NMDAR could treat depression has made no sense.

Zanos and colleagues’ initial experiments placed doubt on an NMDAR-mediated mechanism of action by ketamine (Fig. 1). The authors compared the effects of two different structural forms, or enantiomers, of ketamine, called (S)- and (R)-ketamine, which are normally administered together. (S)-Ketamine is

three to four times better at blocking NMDAR than (R)-ketamine¹³, and so is predicted to be the better antidepressant under the NMDAR-inhibition model. However, the authors found that (R)-ketamine was several times more efficient at reducing depression-like behaviours in mouse models of depression. Furthermore, they confirmed¹⁴ that an even more potent NMDAR inhibitor, which binds to the same site as ketamine, fails to produce sustained antidepressant-like effects.

So what could be responsible for the effects of ketamine treatment? The first hint came from comparing the drug’s activity in male and female mice. Zanos *et al.* confirmed a previous observation¹⁵ that a lower dose of ketamine is needed to reduce depression-like behaviours in females than in males. This could not be explained by different levels of ketamine in the brain. However, the authors

found that levels of the ketamine metabolite hydroxynorketamine (HNK) were several-fold higher in the brains of females than males after the animals were given the same dose of the drug. Reducing the metabolism of ketamine to HNK reduced the effectiveness of ketamine towards depression-related behaviours in mice. Moreover, treating animals with HNK produced the same rapid and sustained antidepressant-like effects seen after treatment with ketamine. As with ketamine, the (R)-enantiomer of HNK had more-potent antidepressant-like effects than the (S)- form. And, importantly, the researchers showed that HNK neither binds to nor inhibits NMDAR.

The finding that the antidepressant effects of ketamine are not mediated through its actions on the NMDAR is a major advance. Nevertheless, it leads to an obvious, unanswered question — what is the molecular target of HNK responsible for these effects? This question should engender much activity by academic scientists, and possibly by large pharmaceutical companies that have been pouring capital into developing NMDAR inhibitors for treating depression. Candidate targets will probably soon emerge.

Although Zanos and colleagues did not identify such a target, they examined the role of another neural receptor protein, AMPAR, which is concentrated at synapses and mediates most neurotransmission in the brain. They found that a drug called NBQX, which reduces AMPAR activity throughout the brain,

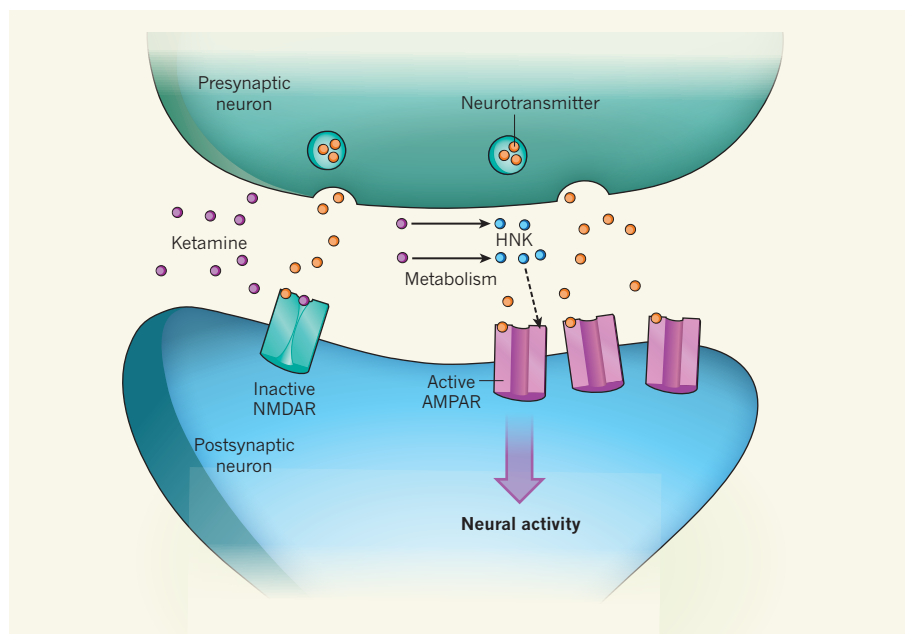


Figure 1 | Metabolite mediator of ketamine. How the drug ketamine exerts its antidepressant effects is unknown, although a common hypothesis states that it acts by binding to the receptor protein NMDAR on postsynaptic neurons, preventing neurotransmitter molecules released by presynaptic neurons from activating NMDAR and so inhibiting signalling processes triggered by the receptor. By contrast, Zanos *et al.*⁷ report that it is a metabolite of ketamine called hydroxynorketamine (HNK) that has antidepressant activity. They provide evidence that HNK, through unknown intermediates, increases the levels of another neuronal receptor protein, AMPAR, at synapses (dashed arrow), enhancing neural activity. But how this produces an antidepressant effect remains unclear.

prevented and even reversed the antidepressant-like effects of ketamine and HNK in mice. It is surprising that a drug that indiscriminately reduces transmission in almost every brain circuit could alter the very specific effects of HNK and ketamine. The authors also show that transient application of HNK produces a long-lasting increase in AMPAR-mediated synaptic transmission (Fig. 1). How this can alleviate depression is not clear, unless HNK acts specifically to modulate the synapses that exhibit reduced function during depression¹⁶. Such a targeted action for HNK remains to be demonstrated.

Finally, Zanos *et al.* show that HNK does not elicit several of the cognitive and motor side effects that have been linked to ketamine. As such, this study represents important progress. Nonetheless, the molecular target

and mechanism of action of HNK remain to be defined. Such advances might further the development of more-specific and effective treatments, allowing people with depression to step out of the darkness of this disorder. ■

Roberto Malinow is in the Department of Neurosciences, School of Medicine, and in the Section of Neurobiology, Division of Biology, University of California, San Diego, La Jolla, California 92093, USA.
e-mail: rmalinow@ucsd.edu

1. Styron, W. *Darkness Visible: A Memoir of Madness* (Random house, 1990).
2. Trevino, K., McClintock, S. M., McDonald Fischer, N., Vora, A. & Husain, M. M. *Ann. Clin. Psychiatry* **26**, 222–232 (2014).
3. Berman, R. M. *et al. Biol. Psychiatry* **47**, 351–354 (2000).

4. Zarate, C. A. Jr *et al. Arch. Gen. Psychiatry* **63**, 856–864 (2006).
5. McGirr, A. *et al. Psychol. Med.* **45**, 693–704 (2015).
6. DiazGranados, N. *et al. J. Clin. Psychiatry* **71**, 1605–1611 (2010).
7. Zanos, P. *et al. Nature* **533**, 481–486 (2016).
8. Anis, N. A., Berry, S. C., Burton, N. R. & Lodge, D. *Br. J. Pharmacol.* **79**, 565–575 (1983).
9. Collingridge, G. L., Kehl, S. J. & McLennan, H. *J. Physiol. (Lond.)* **334**, 33–46 (1983).
10. Bliss, T. V. P. & Lomo, T. *J. Physiol. (Lond.)* **232**, 331–356 (1973).
11. Morris, R. G., Anderson, E., Lynch, G. S. & Baudry, M. *Nature* **319**, 774–776 (1986).
12. Nabavi, S. *et al. Nature* **511**, 348–352 (2014).
13. Ebert, B., Mikkelsen, S., Thorkildsen, C. & Borgbjerg, F. M. *Eur. J. Pharmacol.* **333**, 99–104 (1997).
14. Autry, A. E. *et al. Nature* **475**, 91–95 (2011).
15. Carrier, N. & Kabbaj, M. *Neuropharmacology* **70**, 27–34 (2013).
16. Li, N. *et al. Science* **329**, 959–964 (2010).

This article was published online on 4 May 2016.

ATMOSPHERIC SCIENCE

Unexpected player in particle formation

Three studies find that a family of organic compounds affects the formation and initial growth of atmospheric aerosol particles in clean air — with implications for our knowledge of the climate effects of aerosols. [SEE LETTERS P.521 & 527](#)

CHRIS CAPPA

Cloud droplets form when water condenses on microscopic aerosol particles¹. A key source of new particles in the atmosphere is nucleation — the formation and growth of molecular clusters, which must then grow about 50 times larger if they are to act as efficient cloud seeds. Sulfuric acid has long been recognized as the key player in particle formation². But two studies in this issue^{3,4}, and another published in *Science*⁵, suggest that molecules called highly oxidized multifunctional organic compounds (HOM compounds) have an under-appreciated role in driving both particle formation and the initial growth of particles, especially in environments largely unaffected by anthropogenic pollution.

Understanding the differences between past and present particle formation and growth rates is crucial in quantifying the aerosol cooling effect⁶, which has offset warming driven by greenhouse gases over the past century, but remains highly uncertain⁷. Atmospheric sulfur emissions are higher today than in pre-industrial times because of increased fossil-fuel combustion⁸, so to understand how particles affected the climate in the past, and how they affect pristine regions of the atmosphere today, it is necessary to characterize particle formation and growth when sulfuric

acid concentrations are low. The latest studies together indicate that HOM compounds are key players.

HOM compounds form when hydrocarbons and other volatile organic compounds (VOCs), emitted into the atmosphere from many natural and anthropogenic sources, react with atmospheric oxidants, such as ozone^{9,10}. They are diverse, containing varying numbers of molecules from a wide range of chemical groups, including alcohols and peroxides. Consequently, their vapour pressures — a property that determines their ability to condense — vary by more than 15 orders of magnitude⁴.

Kirkby *et al.*³ (page 521) investigated how effective HOM compounds are at producing new particles with diameters larger than 1.7 nanometres at low sulfuric acid concentrations, whereas Tröstl *et al.*⁴ (page 527) determined the role of HOM compounds in the particles' subsequent growth (for particles starting at about 2 nm in diameter and increasing to about 20 nm). Both studies were performed in the laboratory, and used a VOC called α -pinene — a molecule emitted by trees and from the ocean — as the source of HOM compounds.

In their study, Kirkby *et al.* demonstrate that HOM compounds can nucleate to form particles without sulfuric acid, and that

the particle-formation rate depends on the presence of Galactic cosmic rays (GCRs). Although previous observations^{11,12} showed that organic compounds can enhance sulfuric acid-driven particle-formation rates, a direct demonstration of particle formation by organics in the absence of sulfuric acid had been elusive. The dependence of the organic-driven particle-formation rate on GCRs provides a potential connection between the magnetic variability of the Sun (which affects the GCR flux to Earth), particles and climate, an association that remains widely debated.

Newly formed nanoparticles grow through condensation. The growth stage is crucial for particles less than 10 nm in diameter, because they are especially prone to being absorbed by larger particles on collision, thus removing potential cloud seeds from the atmosphere. Nanoparticle-growth rates increase with diameter¹³, perhaps because of condensation of organic compounds¹⁴, but disentangling the controlling factors has been challenging.

Tröstl *et al.* show that the Kelvin effect — in which the volatility of liquids increases when their interface with the surrounding vapour is curved — rapidly decreases at nanoparticle surfaces as the particles grow. This allows increasingly efficient condensation of HOM compounds that have progressively higher (but always very low) volatilities as the particles grow. Importantly, the accelerating growth rates directly result from the fact that HOM compounds have a distribution of volatilities.

In complement to the two laboratory studies, Bianchi *et al.*⁵ used field observations made at the Jungfraujoch research station in Switzerland (Fig. 1) to show that, when sulfuric acid concentrations are low, particle formation and accelerating growth are indeed efficient only when concentrations of HOM compounds are sufficiently large. Although the observed particle-formation rates are in reasonable agreement with Kirkby and colleagues' results, Bianchi *et al.* were unable



Figure 1 | The Jungfraujoch research station in Switzerland. Bianchi *et al.*⁵ report that highly oxidized organic molecules have a key role in the formation and growth of aerosol particles in the atmosphere, based on measurements taken at Jungfraujoch. Their findings are supported by two laboratory studies^{3,4}. The research station is the small building on top of the grey outcrop of rock, framed by blue sky, in the centre of the landscape.

FOTOVOYAGER/GETTY

to reproduce the observed acceleration in growth rates for particles less than 10 nm in diameter using a mathematical model; by contrast, Tröstl and co-workers were able to model the acceleration in growth rates observed in their study. The authors also found that the HOM compounds at Jungfraujoch were probably anthropogenic in origin, rather than biogenic, suggesting that many VOCs are HOM-compound precursors. Regardless of origin, it seems that the contribution of HOM compounds to nucleation and growth increases as sulfuric acid concentrations fall.

Tröstl and colleagues used their experimental results to constrain simulations made using a global aerosol model. These simulations indicate that the concentration of efficient cloud seeds in the modern atmosphere increases substantially when HOM compounds are included in nanoparticle growth — consistent with the findings of other groups (see ref. 15, for example). Previously reported simulations⁶ indicated that our understanding of how aerosols affect clouds and climate is limited largely by uncertainties in the natural sulfur cycle, but they considered only particle formation induced by sulfuric acid. The latest results suggest that this view must be reassessed, and that uncertainties stemming from the natural VOC cycle are probably larger than was thought.

One challenge in developing robust predictions from these three studies is that HOM compounds were defined as only those that can be detected using a particular type of mass spectrometer. But the detectabilities of HOM

compounds of different compositions and volatilities are not fully established. Although the total concentration of HOM compounds correlated with particle-formation rates and growth rates in both laboratory studies^{3,4}, Tröstl and colleagues' results show that HOM-compound identity cannot be neglected.

Furthermore, different VOCs are not equally efficient at producing HOM compounds¹⁰. The empirical, laboratory-derived relationships were determined only for α -pinene, and so it remains to be seen whether they are generally robust; the mismatch between Bianchi and colleagues' field observations and the laboratory-based predictions suggests that more work is needed. Another issue is that the molecular forces that determine the stability of clusters made purely from HOM compounds are unknown. Nonetheless, the three papers provide a solid foundation for understanding the effects of atmospheric organic compounds on particle abundances in the past, present and future.

The current studies focus on the role of HOM-compound vapours in particle formation and the initial stages of growth. But aerosol particles are microreactors in which chemical reactions occur after, or even during, condensation. This transforms particle compositions¹⁶ and can therefore influence the overall life cycle and climate impacts of particles by altering their volatilities^{17,18}, interactions with water¹⁹ and reactivity²⁰. A better understanding is needed of how the compositions of HOM compounds in vapour (which were measured in these studies) affect

the molecular composition of particles, to establish the full life cycle of aerosols and their effects on the atmosphere. ■

Chris Cappa is in the Department of Civil and Environmental Engineering, University of California, Davis, Davis, California 95616, USA.

e-mail: cdcappa@ucdavis.edu

- Farmer, D. K., Cappa, C. D. & Kreidenweis, S. M. **115**, 4199–4217 (2015).
- Ball, S. M., Hanson, D. R., Eisele, F. L. & McMurry, P. H. *J. Geophys. Res. Atmos.* **104**, 23709–23718 (1999).
- Kirkby, J. *et al. Nature* **533**, 521–526 (2016).
- Tröstl, J. *et al. Nature* **533**, 527–531 (2016).
- Bianchi, F. *et al. Science* <http://dx.doi.org/10.1126/science.aad5456> (2016).
- Carlsaw, K. S. *et al. Nature* **503**, 67–71 (2013).
- IPCC. *Climate Change 2013: The Physical Science Basis. Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. (Cambridge Univ. Press, 2013).
- McConnell, J. R. *et al. Science* **317**, 1381–1384 (2007).
- Ehn, M. *et al. Nature* **506**, 476–479 (2014).
- Jokinen, T. *et al. Proc. Natl Acad. Sci. USA* **112**, 7123–7128 (2015).
- Riccobono, F. *et al. Science* **344**, 717–721 (2014).
- Zhang, R. *et al. Science* **304**, 1487–1490 (2004).
- Kulmala, M. *et al. Science* **339**, 943–946 (2013).
- Riipinen, I. *et al. Nature Geosci.* **5**, 453–458 (2012).
- D'Andrea, S. D. *et al. Atmos. Chem. Phys.* **13**, 11519–11534 (2013).
- Kroll, J. H. & Seinfeld, J. H. *Atmos. Environ.* **42**, 3593–3624 (2008).
- Lopez-Hilfiker, F. D. *et al. Atmos. Chem. Phys.* **15**, 7765–7776 (2015).
- Kolesar, K. R., Chen, C., Johnson, D. & Cappa, C. D. *Atmos. Chem. Phys.* **15**, 9327–9343 (2015).
- Ruehl, C. R., Davies, J. F. & Wilson, K. R. *Science* **351**, 1447–1450 (2016).
- Kroll, J. H., Lim, C. Y., Kessler, S. H. & Wilson, K. R. *J. Phys. Chem. A* **119**, 10767–10783 (2015).

NMDAR inhibition-independent antidepressant actions of ketamine metabolites

Panos Zanos¹, Ruin Moaddel², Patrick J. Morris³, Polymnia Georgiou¹, Jonathan Fischell⁴, Greg I. Elmer^{1,5,6}, Manickavasagam Alkondon⁷, Peixiong Yuan⁸, Heather J. Pribut¹, Nagendra S. Singh², Katina S. S. Dossou², Yuhong Fang³, Xi-Ping Huang⁹, Cheryl L. Mayo⁶, Irving W. Wainer^{2†}, Edson X. Albuquerque^{5,7,10}, Scott M. Thompson^{1,4}, Craig J. Thomas³, Carlos A. Zarate Jr⁸ & Todd D. Gould^{1,5,11}

Major depressive disorder affects around 16 per cent of the world population at some point in their lives. Despite the availability of numerous monoaminergic-based antidepressants, most patients require several weeks, if not months, to respond to these treatments, and many patients never attain sustained remission of their symptoms. The non-competitive, glutamatergic NMDAR (*N*-methyl-D-aspartate receptor) antagonist (*R,S*)-ketamine exerts rapid and sustained antidepressant effects after a single dose in patients with depression, but its use is associated with undesirable side effects. Here we show that the metabolism of (*R,S*)-ketamine to (2*S*,6*S*;2*R*,6*R*)-hydroxynorketamine (HNK) is essential for its antidepressant effects, and that the (2*R*,6*R*)-HNK enantiomer exerts behavioural, electroencephalographic, electrophysiological and cellular antidepressant-related actions in mice. These antidepressant actions are independent of NMDAR inhibition but involve early and sustained activation of AMPARs (α -amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors). We also establish that (2*R*,6*R*)-HNK lacks ketamine-related side effects. Our data implicate a novel mechanism underlying the antidepressant properties of (*R,S*)-ketamine and have relevance for the development of next-generation, rapid-acting antidepressants.

Major depressive disorder is common, affecting about 16% of the world population at some point in their lives, and is associated with serious health and socioeconomic consequences^{1,2}. Current pharmacotherapies, including monoaminergic-acting antidepressants, require prolonged administration (weeks if not months) for clinical improvement. This lag time, as well as a high non-response rate, emphasizes the need for better antidepressant medications³. The non-competitive, glutamatergic NMDAR antagonist (*R,S*)-ketamine (ketamine) has demonstrated rapid and robust efficacy as an antidepressant by improving core depressive symptoms including depressed mood, anhedonia, and suicidal thoughts in treatment-refractory unipolar and bipolar depressed patients when administered at sub-anaesthetic doses^{4–8}. Remarkably, these actions are observed within hours after a single administration, and persist on average for 1 week. While discovery of the clinical antidepressant efficacy of ketamine for the treatment of depression has elicited tremendous excitement in the field, its potential for widespread clinical use is limited owing to its abuse liability and capacity to produce dissociative effects even when administered at low doses⁹. There are also unanswered questions about how ketamine works as an antidepressant, which is typically assumed to depend on direct NMDAR inhibition. However, the results of human treatment trials indicate that alternative NMDAR antagonists lack the robust, rapid and/or sustained antidepressant properties of ketamine¹⁰.

Role of NMDAR inhibition in ketamine action

We compared the antidepressant-like effects of ketamine and the classical tricyclic antidepressant desipramine in the mouse forced-swim test (FST) at 1 h (acute) and 24 h (sustained) after administration (Fig. 1a). A 10 mg kg^{−1} dose of ketamine resulted in acute and long-lasting dose-dependent antidepressant effects in the FST, whereas desipramine decreased immobility time only 1 h after injection. To date, most studies assessing the antidepressant effects of ketamine are based on a commonly accepted view that ketamine and its *N*-demethylated metabolite (*R,S*)-norketamine are the active agents, the clinical effects of which are due to inhibition of the NMDAR. Additional metabolites (Extended Data Figs 1 and 2a, b) are considered clinically inactive since they do not induce anaesthesia¹¹. To determine whether NMDAR inhibition is the main mechanism underlying the antidepressant effects of ketamine, we assessed the effects of the (*S*)- and (*R*)-ketamine enantiomers in the FST (Fig. 1b), novelty-suppressed feeding (NSF) test (Fig. 1c) and learned helplessness test (Fig. 1d). While the NMDAR hypothesis of ketamine action would predict greater efficacy of (*S*)-ketamine since it is a ~3–4-fold more potent inhibitor of the NMDAR than (*R*)-ketamine^{12,13}, our results, in accordance with a recent report¹⁴, demonstrate a greater potency of (*R*)-ketamine in all three antidepressant-predictive tasks. Notably, this antidepressant effect does not result from higher brain levels of (*R*)-ketamine than of (*S*)-ketamine (Extended Data Fig. 2c–e). Moreover, in contrast to ketamine, we

¹Department of Psychiatry, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA. ²Biomedical Research Center, National Institute on Aging, National Institutes of Health, Baltimore, Maryland 21224, USA. ³Division of Preclinical Innovation, National Center for Advancing Translational Sciences, National Institutes of Health, Rockville, Maryland 20850, USA.

⁴Department of Physiology, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA. ⁵Department of Pharmacology, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA. ⁶Maryland Psychiatric Research Center, University of Maryland School of Medicine, Baltimore, Maryland 21228, USA. ⁷Department of Epidemiology and Public Health, Division of Translational Toxicology, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA. ⁸Experimental Therapeutics and Pathophysiology Branch, Intramural Research Program, National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland 20892, USA. ⁹NIMH Psychoactive Drug Screening Program, Department of Pharmacology and Division of Chemical Biology and Medicinal Chemistry, University of North Carolina Chapel Hill Medical School, Chapel Hill, North Carolina 27516, USA. ¹⁰Department of Medicine, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA. ¹¹Department of Anatomy and Neurobiology, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA.

[†]Present address: Mitchell Woods Pharmaceuticals, Shelton, Connecticut 06484, USA.

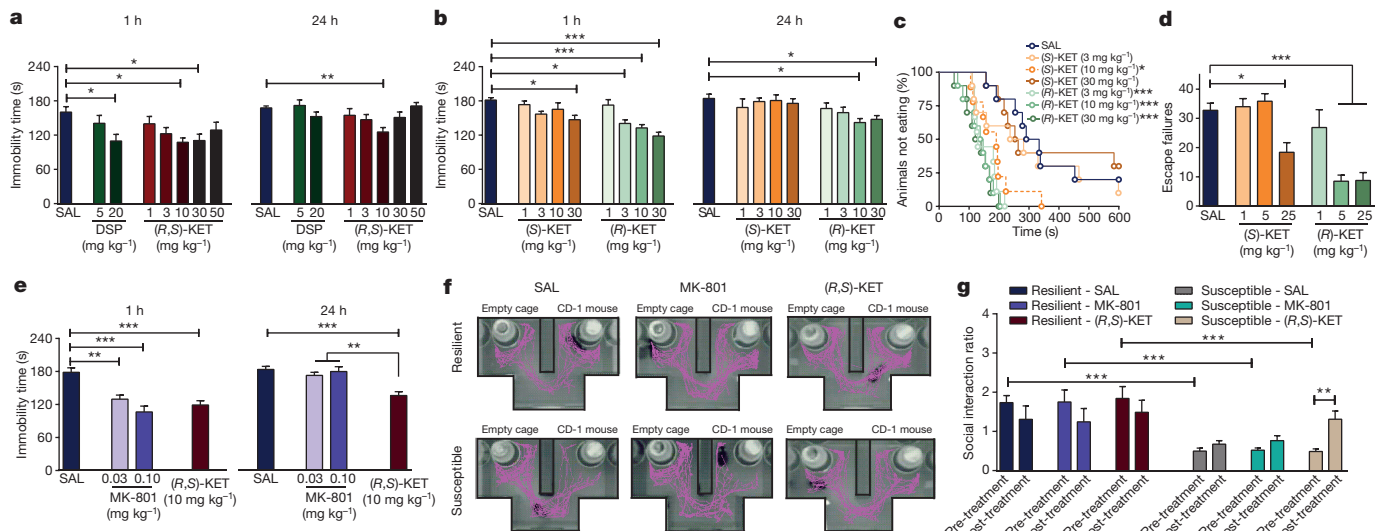


Figure 1 | NMDAR inhibition is not sufficient for the antidepressant actions of ketamine. **a**, Antidepressant-like responses of (R,S)-ketamine (KET) and desipramine (DSP) in the forced-swim test (FST) 1- and 24-h after treatment. SAL, saline. **b–d**, Compared to (S)-KET, (R)-KET showed greater and longer-lasting antidepressant-like effects in the FST (**b**), novelty-suppressed feeding (NSF) test (**c**) and learned helplessness test (**d**).

demonstrated that the NMDAR antagonist MK-801, which binds at the same receptor site as ketamine, does not exert sustained (24 h) antidepressant-like effects in the FST (Fig. 1e; see also refs 15, 16), or reverse social interaction deficits induced by chronic social defeat stress (Fig. 1g and Extended Data Fig. 3). These findings indicate a probable NMDAR inhibition-independent mechanism underlying the antidepressant responses of ketamine.

Antidepressant actions of ketamine metabolites

Ketamine is stereoselectively metabolised into a broad array of metabolites, including norketamine, hydroxyketamines, dehydronorketamine and the HNKs^{17,18} (Fig. 2a and Extended Data Fig. 1). After ketamine administration, (2S,6S;2R,6R)-HNK is the major HNK metabolite found in the plasma and brain of mice (Extended Data Fig. 2a, b), and

e–g, The alternative NMDAR antagonist MK-801 did not elicit 24-h antidepressant actions in the FST (**e**), and did not reverse social avoidance induced by chronic social defeat stress (**f**, **g**), where purple lines represent the video-tracked movements of mice (**f**). Data are mean \pm s.e.m. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ (see Supplementary Table 1 for statistical analyses and n numbers).

plasma of humans¹⁹. Similar to previous evidence revealing enhanced ketamine antidepressant responses in female rodents compared to males^{20,21}, we observed greater antidepressant potency of ketamine in female mice in the FST (Fig. 2b), which was not associated with sex differences in ketamine-induced hyperlocomotion (probably mediated by NMDAR inhibition²²; Extended Data Fig. 4a, b). To investigate whether these sex-dependent antidepressant differences are explained by a different pharmacokinetic profile of ketamine in males versus females, we measured the levels of ketamine and its metabolites in the brains of mice after ketamine administration. While equivalent levels of ketamine and norketamine were found, (2S,6S;2R,6R)-HNK was approximately three-fold higher in the brains of female mice compared to males (Fig. 2c–e), suggesting a role of (2S,6S;2R,6R)-HNK in the antidepressant effects of ketamine. To directly determine

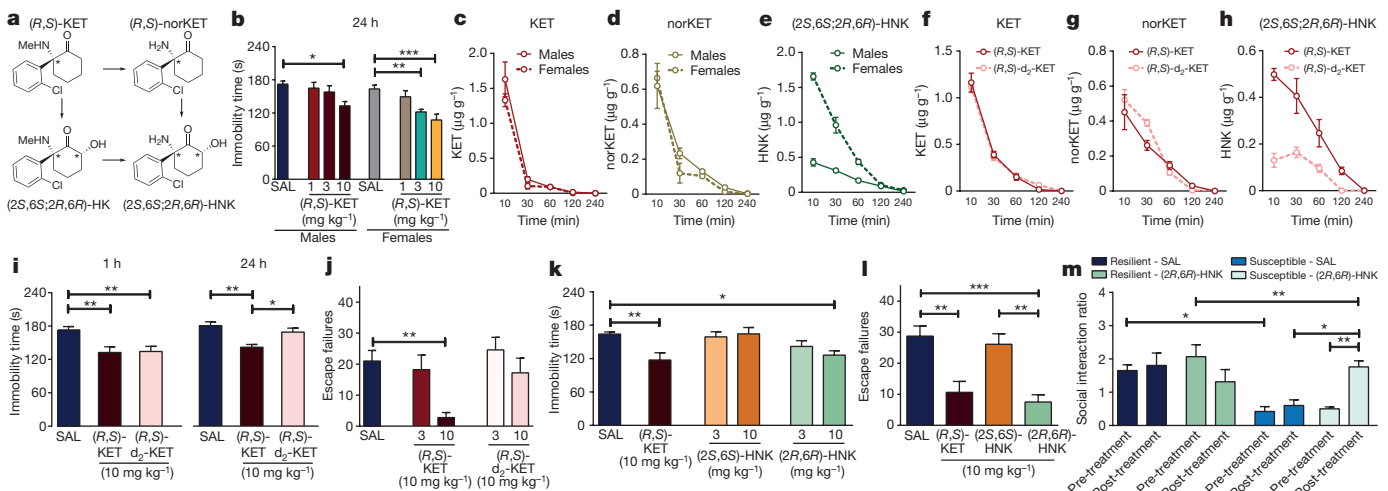


Figure 2 | Metabolism of ketamine to (2R,6R)-HNK is necessary and sufficient to exert antidepressant actions. **a**, Simplified diagram of (R,S)-KET metabolism. **b–e**, Greater antidepressant-like actions of ketamine in female mice compared to males in the FST (**b**) are associated with higher brain levels of (2S,6S;2R,6R)-HNK (**e**), but not KET (**c**) or norketamine (norKET) (**d**). **f–h**, Brain levels of KET (**f**), norKET (**g**) and (2S,6S;2R,6R)-HNK (**h**) after administration of (R,S)-KET and 6,6-dideuteroketamine ((R,S)-d₂-KET). **i, j**, Effects of (R,S)-KET and

(R,S)-d₂-KET in the 1-h and 24-h FST (**i**) and the learned helplessness test (**j**). **k, l**, Compared to (2S,6S)-HNK, (2R,6R)-HNK manifested greater potency and longer-lasting antidepressant-like effects in the FST (**k**) and learned helplessness test (**l**). **m**, (2R,6R)-HNK reversed chronic social defeat-induced social interaction deficits. Data are mean \pm s.e.m. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ (see Supplementary Table 1 for statistical analyses and n numbers).

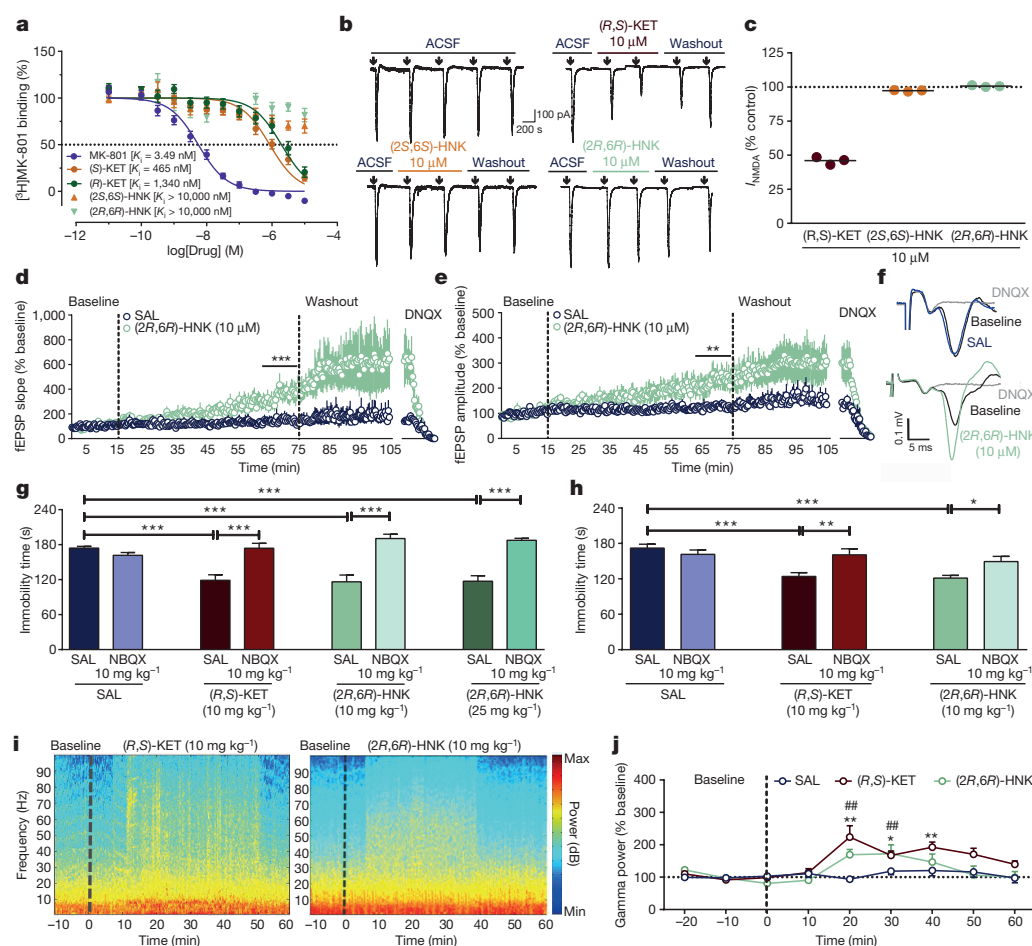


Figure 3 | Role of NMDA and AMPA glutamate receptors in the acute antidepressant effects of (2R,6R)-HNK. **a**, (2R,6R)-HNK does not displace [³H]MK-801 binding. **b**, **c**, (R,S)-KET inhibited, but (2S,6S)-HNK and (2R,6R)-HNK did not inhibit currents evoked by application of NMDA to stratum radiatum interneurons in rat hippocampal slices (**b**), quantified as percentage inhibition (I_{NMDA} ; **c**). Arrows indicate 30-s agonist pulse. ACSF, artificial cerebrospinal fluid. **d**, **e**, Normalized fEPSP slope (**d**) and amplitude (**e**) from stimulation of the Schaffer collateral pathway in rat hippocampal slices. **f**, Representative field-potential traces in the same hippocampal slice before (baseline) and 60 min after application of SAL or (2R,6R)-HNK. **g**, **h**, Pre-treatment with the AMPAR inhibitor NBQX 10 min before (R,S)-KET or (2R,6R)-HNK prevented their antidepressant-like actions in the 1-h (**g**) or 24-h (**h**) FST. **i**, Representative qEEG spectrograms for 10-min before (baseline) and 1-h after administration of (R,S)-ketamine or (2R,6R)-HNK (indicated by a dashed line). **j**, Normalized gamma power changes after administration of (R,S)-KET, (2R,6R)-HNK or vehicle (SAL). Data are mean \pm s.e.m. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$; in **j**, * denotes (R,S)-KET, # denotes (2R,6R)-HNK (see Supplementary Table 1 for statistical analyses and n numbers).

whether metabolism of ketamine to (2S,6S;2R,6R)-HNK is required for its antidepressant actions, we deuterated ketamine at the C6 position (6,6-dideuteroketamine; (R,S)-d₂-KET, Extended Data Fig. 2f). This alteration would not change the pharmacological properties of unmetabolized ketamine, but may change the relative rate of metabolism²³. Indeed, 6,6-dideuteroketamine did not change NMDAR binding affinity (Extended Data Fig. 2g), or NMDAR-mediated hyperlocomotion (Extended Data Fig. 4c, d), but robustly hindered its metabolism to (2S,6S;2R,6R)-HNK, without changing the ketamine levels in the brain (Fig. 2f–h). Unlike ketamine, administration of 6,6-dideuteroketamine did not induce antidepressant actions in the FST (Fig. 2i) or learned helplessness test (Fig. 2j) 24 h after administration, indicating a role of (2S,6S;2R,6R)-HNK in the sustained antidepressant effects. Notably, published human data reveal a positive correlation between the antidepressant responses of ketamine and plasma (2S,6S;2R,6R)-HNK metabolite levels¹⁹.

To determine whether (2S,6S)-HNK or (2R,6R)-HNK exert antidepressant effects independently of ketamine administration, we compared their behavioural effects in the 24-h (sustained) FST and learned helplessness test. We observed more potent antidepressant effects after administration of the (2R,6R)-HNK metabolite (Fig. 2k, l), which is exclusively derived from (R)-ketamine, and thus consistent with the greater antidepressant actions of (R)-ketamine relative to (S)-ketamine (Fig. 1b–d). Moreover, (2R,6R)-HNK resulted in a dose-dependent antidepressant action in the learned helplessness test, FST and NSF test (Extended Data Fig. 5a, c, f). We note that (2S,6S)-HNK also exerts antidepressant actions at higher doses (Extended Data Fig. 5b, d). The greater antidepressant effects of (2R,6R)-HNK do not result from higher brain levels of the drug compared to (2S,6S)-HNK (Extended Data Fig. 5e). Similar to ketamine, a single (2R,6R)-HNK administration

induced persistent antidepressant effects in the FST, lasting for at least 3 days (Extended Data Fig. 5g). A single (2R,6R)-HNK administration also reversed chronic corticosterone-induced anhedonia assessed with the sucrose preference and female urine sniffing behavioural tasks (Extended Data Fig. 5h, i), as well as social avoidance induced by chronic social defeat stress (Fig. 2m; Extended Data Fig. 5j, k).

(2R,6R)-HNK effects on glutamate receptors

A prominent hypothesis for the mechanism of action of ketamine is that it acts via direct inhibition of NMDARs localized to interneurons. This is suggested to lead to disinhibition of glutamatergic neurons, which receive input from interneurons, and a resultant rapid increase in glutamate synaptic transmission in mood-relevant brain regions²⁴. However, in contrast to ketamine, (2R,6R)-HNK does not displace [³H]MK-801 binding to the NMDAR *in vitro* (Fig. 3a; also see ref. 12) and does not functionally inhibit NMDARs localized to stratum radiatum interneurons in hippocampal slices (Fig. 3b, c). Instead, (2R,6R)-HNK induced a robust increase in AMPAR-mediated excitatory post-synaptic potentials (EPSPs) recorded from the CA1 region of hippocampal slices after stimulation of Schaffer collateral axons, which was sustained after washout of the drug (Fig. 3d–f). (2R,6R)-HNK also increased the frequency and amplitude of AMPAR-mediated excitatory postsynaptic currents (EPSCs) recorded from CA1 stratum radiatum interneurons (Extended Data Fig. 6a–j), which receive glutamatergic inputs from the Schaffer collaterals. To test the extent to which the antidepressant effect of (2R,6R)-HNK depends on AMPAR activation *in vivo*, mice were pre-treated with the AMPAR antagonist 2, 3-dihydroxy-6-nitro-7-sulfamoyl-benzof[*q*]quinoxaline-2, 3-dione (NBQX) 10 min before treatment with ketamine or (2R,6R)-HNK. Mice were then assessed in the FST 1 or 24 h after the treatment.

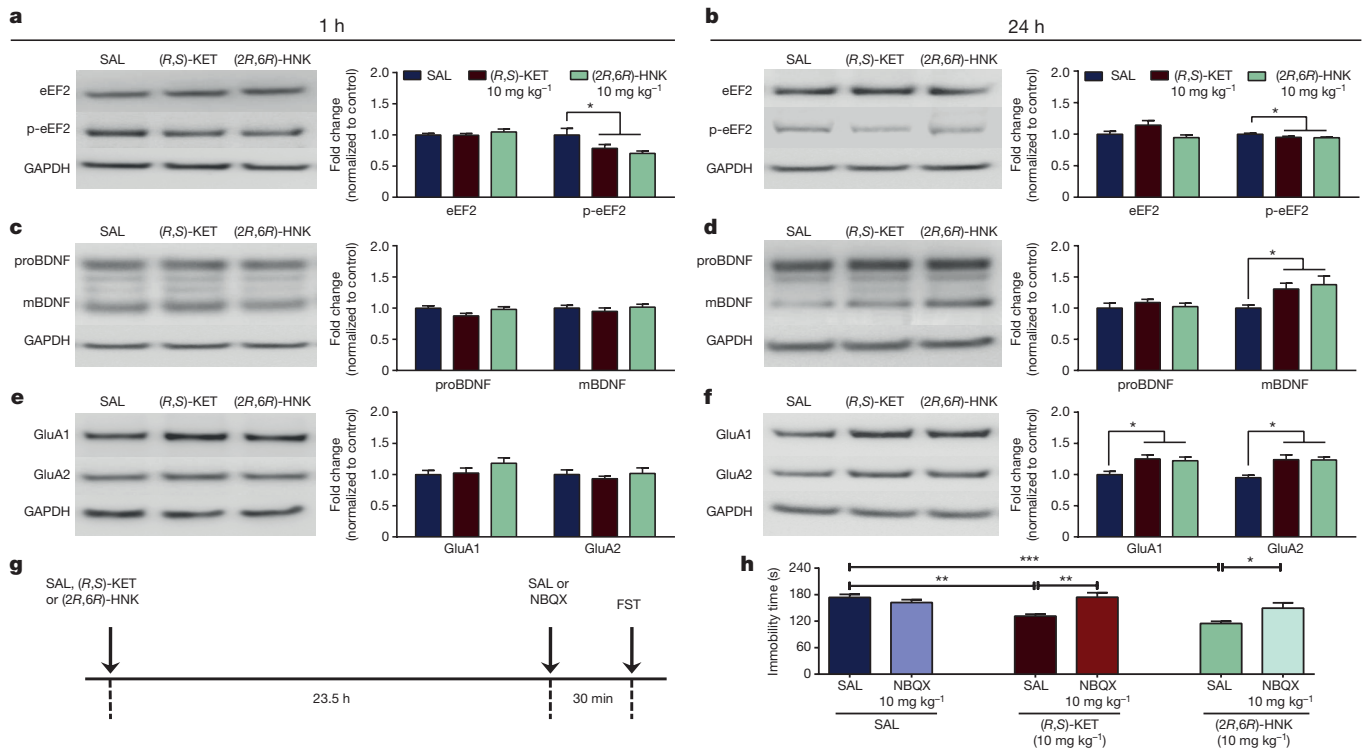


Figure 4 | Role of AMPARs in the sustained antidepressant effects of (2R,6R)-HNK. **a–h**, Protein and protein phosphorylation levels from hippocampal synaptoneurosome fractions. **a, b**, A single administration of (R,S)-KET or (2R,6R)-HNK decreased phosphorylation of eEF2 (p-eEF2), 1 h (**a**) and 24 h (**b**) after injection. **c, d**, Although administration of (2R,6R)-HNK or (R,S)-KET did not alter the levels of proBDNF or mature BDNF (mBDNF) 1 h after injection (**c**), it increased mBDNF levels 24 h after treatment (**d**). **e, f**, (R,S)-KET and (2R,6R)-HNK did not change

Similar NBQX treatment has previously been shown to prevent the antidepressant actions of ketamine, without affecting other behaviours in rodents^{15,25–27}. Treatment with NBQX, prior to (2R,6R)-HNK, prevented both the 1-h and 24-h antidepressant effects of (2R,6R)-HNK (Fig. 3g, h), indicating that its antidepressant actions require the acute activation of AMPARs.

A non-invasive method used to assess ketamine-activated circuitry in both humans and rodents is the quantitative electroencephalography (qEEG) measurement of gamma-band power, which is dependent on activation of fast ionotropic excitatory receptors, including AMPARs^{28–30}. We show that, similar to ketamine, (2R,6R)-HNK administration acutely increases gamma power measured via surface electrodes *in vivo* (Fig. 3i, j), independent of locomotor activity changes, and without altering alpha, beta, delta or theta oscillations (Extended Data Fig. 7a–e). Importantly, pre-treatment with NBQX prevented (2R,6R)-HNK-induced increases in gamma power, thus further implicating AMPARs in the (2R,6R)-HNK mechanism of action (Extended Data Fig. 7f–k), and validating a potential human translational biomarker of the central nervous system response to (2R,6R)-HNK.

Evidence indicates that mammalian target of rapamycin (mTOR) signalling²⁵, protein synthesis through eukaryotic translation elongation factor 2 (eEF2) dephosphorylation¹⁶, as well as brain-derived neurotrophic factor (BDNF) increases^{16,31}, underlie the antidepressant responses of ketamine. We examined whether administration of (2R,6R)-HNK affects phosphorylation of mTOR (Ser2448) and eEF2 (Thr56), or BDNF levels in synaptoneurosome fractions of the hippocampus and prefrontal cortex. No differences were observed in mTOR phosphorylation after administration of ketamine or (2R,6R)-HNK in the hippocampus or prefrontal cortex of mice (Extended Data Fig. 8a–d). However, ketamine induced a decrease in eEF2 phosphorylation in the hippocampus 1 h and 24 h after injection, and increased hippocampal

GluA1 and GluA2 levels at 1 h after treatment (**e**), but did increase levels 24 h after injection (**f**). **g, h**, Administration of the AMPAR inhibitor NBQX 30 min before the 24-h FST prevented the antidepressant effects of both (R,S)-KET and (2R,6R)-HNK administered 23.5 h before NBQX. Data are mean \pm s.e.m. Images cropped; see Supplementary Fig. 1 for complete blot images. GAPDH was used as a loading control. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ (see Supplementary Table 1 for statistical analyses and n numbers).

BDNF at 24 h (Fig. 4a–d). These changes did not occur in the prefrontal cortex (Extended Data Fig. 8e–h), but were recapitulated by (2R,6R)-HNK administration (Fig. 4a–d), and may be partially responsible for its sustained antidepressant actions.

It is noteworthy that (2R,6R)-HNK resulted in antidepressant actions (Fig. 2k–m; Extended Data Fig. 5) at time points (for example, 24 h) past when its brain concentrations are below detectable levels (for example, 2 h; Extended Data Fig. 5e). Synaptic plasticity changes involving AMPARs are thought to underlie such long-term antidepressant actions of ketamine^{24,27}. Here we show that while neither ketamine nor (2R,6R)-HNK administration altered the levels of AMPAR subunits GluA1 and GluA2 in hippocampal synaptoneurosome 1 h after treatment (Fig. 4e), they both increased GluA1 and GluA2 levels 24 h after treatment in mouse hippocampal (Fig. 4f), but not prefrontal cortex synaptoneurosome (Extended Data Fig. 8i, j). Consistent with an increase in synaptic AMPARs being involved in the sustained, 24-h, antidepressant actions, administration of NBQX 30 min prior to the 24-h FST (23.5 h after antidepressant treatment; see timeline Fig. 4g) prevented the antidepressant actions of both ketamine and (2R,6R)-HNK (Fig. 4h). These findings implicate an AMPAR-mediated maintenance of synaptic potentiation to underlie the sustained antidepressant effects of (2R,6R)-HNK.

(2R,6R)-HNK lacks ketamine-related side effects

Ketamine has abuse potential, as well as sensory-dissociation properties and other side effects, which limit its potential widespread use for the treatment of depression⁹. While administration of ketamine (Extended Data Fig. 4a–d) and (2S,6S)-HNK (Fig. 5a) were associated with increased locomotor activity and motor incoordination (Fig. 5c, d), (2R,6R)-HNK did not induce any significant changes in locomotion, and did not affect coordination as measured by the accelerating rotarod test (Fig. 5b, d). We show that unlike ketamine, (2R,6R)-HNK

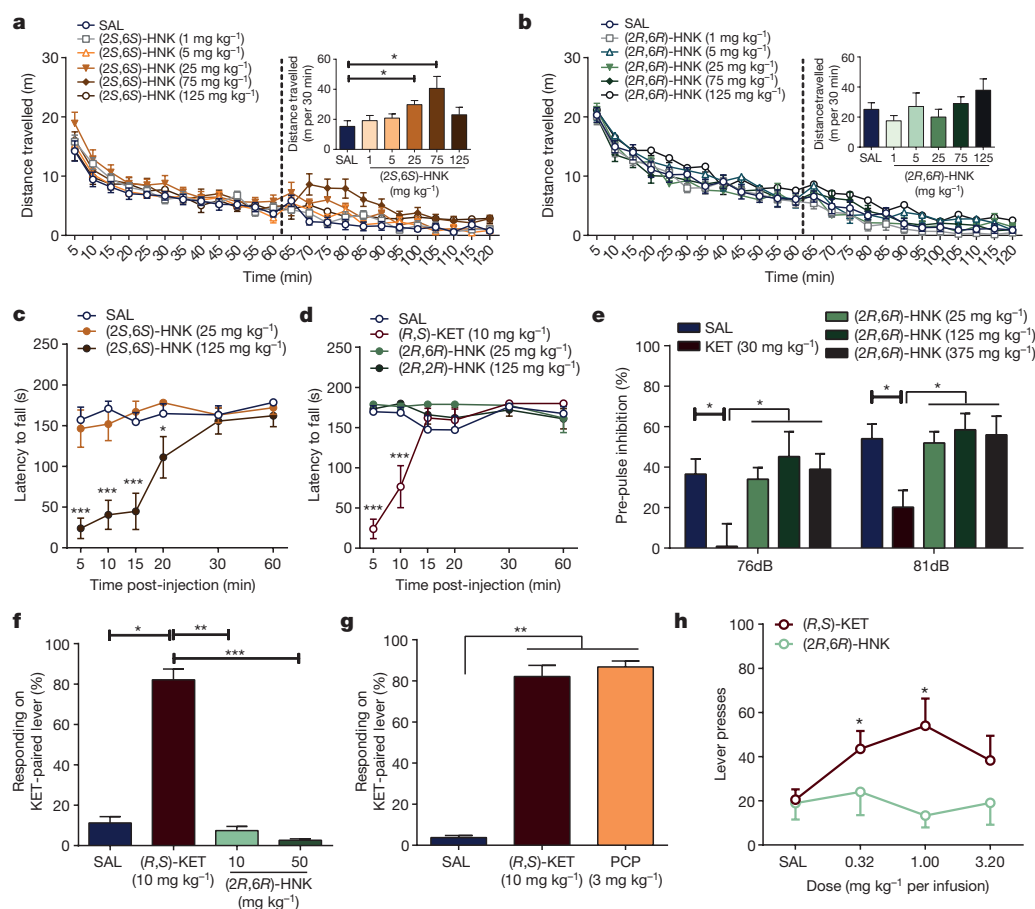


Figure 5 | (2R,6R)-HNK lacks side effects of ketamine.

a, b, After recording baseline activity for 1 h, mice received drug (dashed line) and locomotor activity was monitored for 1 h. Administration of (2S,6S)-HNK dose-dependently changed locomotor activity (**a**), whereas administration of (2R,6R)-HNK did not (**b**). **c, d**, (2S,6S)-HNK (**c**) but not (2R,6R)-HNK (**d**) induced motor in-coordination in the rotarod. **e–h**, Unlike (R,S)-KET, (2R,6R)-HNK administration did not induce pre-pulse inhibition deficits (**e**), (R,S)-KET-associated discriminative stimulus (**f, g**), or self-administration (**h**). Data are mean \pm s.e.m. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ (see Supplementary Table 1 for statistical analyses and n numbers).

administration, even at high doses (375 mg kg⁻¹), did not affect sensory gating as assessed with pre-pulse inhibition (Fig. 5e) or startle amplitude (Extended Data Fig. 9a). Non-competitive NMDAR antagonists, including ketamine and phencyclidine, produce discriminative stimulus effects in drug discrimination protocols and manifest cross-drug substitution profiles at an antidepressant-relevant dose range³². In ketamine-trained mice, (2R,6R)-HNK administration did not produce ketamine-related discrimination responses, whereas phencyclidine (PCP) did (Fig. 5f, g), without either of these drugs changing overall lever pressing response rates (Extended Data Fig. 9b, c). These findings further support a non-NMDAR mechanism for (2R,6R)-HNK action including interoceptive effects, unlike the abused drugs ketamine and PCP. Since drug discrimination does not independently predict abuse potential per se, we further assessed the effects of ketamine and (2R,6R)-HNK in an intravenous drug self-administration model, classically used for the evaluation of abuse/addiction liability. Intravenous ketamine was readily self-administered and resulted in a significant increase in drug intake (Fig. 5h; Extended Data Fig. 9d). By contrast, mice did not self-administer pharmacologically relevant doses of (2R,6R)-HNK under the same conditions (Fig. 5h; Extended Data Fig. 9d). Overall, (2R,6R)-HNK administration revealed an innocuous side-effect profile compared to ketamine.

Discussion

Our data provide new evidence explaining the unique antidepressant effects of ketamine and implicate an NMDAR inhibition-independent mechanism. These findings reveal that production of a distinct metabolite of ketamine is necessary and sufficient to produce the ketamine antidepressant actions. Overall, our data indicate that administration of (2R,6R)-HNK induces an acute increase in glutamatergic signaling (as supported by our EPSP, EPSC and qEEG measurements), followed by a long-term adaptation involving the upregulation of

synaptic AMPARs, as evidenced by an increase in GluA1 and GluA2 in hippocampal synapses. This is supported by the finding that NBQX reverses both the acute (delivered before (2R,6R)-HNK) (Fig. 3g, h) and sustained (delivered after (2R,6R)-HNK; Fig. 4g, h) antidepressant actions of (2R,6R)-HNK. Considering the lack of side effects, and the favourable physiochemical properties of HNKs³³, these findings have relevance for the development of next-generation, rapid-acting antidepressants.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 October 2015; accepted 12 April 2016.

Published online 4 May 2016.

- Kessler, R.C. *et al.* The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *JAMA* **289**, 3095–3105 (2003).
- Trivedi, M. H. *et al.* Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. *Am. J. Psychiatry* **163**, 28–40 (2006).
- Rush, A. J. *et al.* Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. *Am. J. Psychiatry* **163**, 1905–1917 (2006).
- Zarate, C. A., Jr *et al.* A randomized trial of an N-methyl-D-aspartate antagonist in treatment-resistant major depression. *Arch. Gen. Psychiatry* **63**, 856–864 (2006).
- Berman, R. M. *et al.* Antidepressant effects of ketamine in depressed patients. *Biol. Psychiatry* **47**, 351–354 (2000).
- Diazgranados, N. *et al.* A randomized add-on trial of an N-methyl-D-aspartate antagonist in treatment-resistant bipolar depression. *Arch. Gen. Psychiatry* **67**, 793–802 (2010).
- Lally, N. *et al.* Anti-anhedonic effect of ketamine and its neural correlates in treatment-resistant bipolar depression. *Transl. Psychiatry* **4**, e469 (2014).
- Murrough, J. W. *et al.* Antidepressant efficacy of ketamine in treatment-resistant major depression: a two-site randomized controlled trial. *Am. J. Psychiatry* **170**, 1134–1142 (2013).
- Morgan, C. J. & Curran, H. V. Ketamine use: a review. *Addiction* **107**, 27–38 (2012).
- Newport, D. J. *et al.* Ketamine and other NMDA antagonists: early clinical trials and possible mechanisms in depression. *Am. J. Psychiatry* **172**, 950–966 (2015).

11. Leung, L. Y. & Baillie, T. A. Comparative pharmacology in the rat of ketamine and its two principal metabolites, norketamine and (Z)-6-hydroxynorketamine. *J. Med. Chem.* **29**, 2396–2399 (1986).
12. Moaddel, R. *et al.* Sub-anesthetic concentrations of (R,S)-ketamine metabolites inhibit acetylcholine-evoked currents in $\alpha 7$ nicotinic acetylcholine receptors. *Eur. J. Pharmacol.* **698**, 228–234 (2013).
13. Ebert, B., Mikkelsen, S., Thorkildsen, C. & Borgbjerg, F. M. Norketamine, the main metabolite of ketamine, is a non-competitive NMDA receptor antagonist in the rat cortex and spinal cord. *Eur. J. Pharmacol.* **333**, 99–104 (1997).
14. Yang, C. *et al.* R-ketamine: a rapid-onset and sustained antidepressant without psychotomimetic side effects. *Transl. Psychiatry* **5**, e632 (2015).
15. Maeng, S. *et al.* Cellular mechanisms underlying the antidepressant effects of ketamine: role of α -amino-3-hydroxy-5-methylisoxazole-4-propionic acid receptors. *Biol. Psychiatry* **63**, 349–352 (2008).
16. Autry, A. E. *et al.* NMDA receptor blockade at rest triggers rapid behavioural antidepressant responses. *Nature* **475**, 91–95 (2011).
17. Desta, Z. *et al.* Stereoselective and regiospecific hydroxylation of ketamine and norketamine. *Xenobiotica* **42**, 1076–1087 (2012).
18. Adams, J. D., Jr, Baillie, T. A. & Trevor, A. J. & Castagnoli, N. Jr. Studies on the biotransformation of ketamine. 1-Identification of metabolites produced *in vitro* from rat liver microsomal preparations. *Biomed. Mass Spectrom.* **8**, 527–538 (1981).
19. Zarate, C. A., Jr *et al.* Relationship of ketamine's plasma metabolites with response, diagnosis, and side effects in major depression. *Biol. Psychiatry* **72**, 331–338 (2012).
20. Carrier, N. & Kabbaj, M. Sex differences in the antidepressant-like effects of ketamine. *Neuropharmacology* **70**, 27–34 (2013).
21. Franceschelli, A., Sens, J., Herchick, S., Thelen, C. & Pitychoutis, P. M. Sex differences in the rapid and the sustained antidepressant-like effects of ketamine in stress-naïve and “depressed” mice exposed to chronic mild stress. *Neuroscience* **290**, 49–60 (2015).
22. Irifune, M., Shimizu, T., Nomoto, M. & Fukuda, T. Involvement of N-methyl-D-aspartate (NMDA) receptors in noncompetitive NMDA receptor antagonist-induced hyperlocomotion in mice. *Pharmacol. Biochem. Behav.* **51**, 291–296 (1995).
23. Gant, T. G. Using deuterium in drug discovery: leaving the label in the drug. *J. Med. Chem.* **57**, 3595–3611 (2014).
24. Duman, R. S., Aghajanian, G. K., Sanacora, G. & Krystal, J. H. Synaptic plasticity and depression: new insights from stress and rapid-acting antidepressants. *Nature Med.* **22**, 238–249 (2016).
25. Li, N. *et al.* mTOR-dependent synapse formation underlies the rapid antidepressant effects of NMDA antagonists. *Science* **329**, 959–964 (2010).
26. Koike, H., Iijima, M. & Chaki, S. Involvement of AMPA receptor in both the rapid and sustained antidepressant-like effects of ketamine in animal models of depression. *Behav. Brain Res.* **224**, 107–111 (2011).
27. Koike, H. & Chaki, S. Requirement of AMPA receptor stimulation for the sustained antidepressant activity of ketamine and LY341495 during the forced swim test in rats. *Behav. Brain Res.* **271**, 111–115 (2014).
28. Whittington, M. A., Traub, R. D., Kopell, N., Ermentrout, B. & Buhl, E. H. Inhibition-based rhythms: experimental and mathematical observations on network dynamics. *Int. J. Psychophysiol.* **38**, 315–336 (2000).
29. Cunningham, M. O., Davies, C. H., Buhl, E. H., Kopell, N. & Whittington, M. A. Gamma oscillations induced by kainate receptor activation in the entorhinal cortex *in vitro*. *J. Neurosci.* **23**, 9761–9769 (2003).
30. Muthukumaraswamy, S. D. *et al.* Evidence that Subanesthetic doses of ketamine cause sustained disruptions of NMDA and AMPA-mediated frontoparietal connectivity in humans. *J. Neurosci.* **35**, 11694–11706 (2015).
31. Lepack, A. E., Fuchikami, M., Dwyer, J. M., Banasr, M. & Duman, R. S. BDNF release is required for the behavioral actions of ketamine. *Int. J. Neuropsychopharmacol.* **18**, pyu033 (2015).
32. De Vry, J. & Jentsch, K. R. Role of the NMDA receptor NR2B subunit in the discriminative stimulus effects of ketamine. *Behav. Pharmacol.* **14**, 229–235 (2003).
33. Moaddel, R. *et al.* The distribution and clearance of (2S,6S)-hydroxynorketamine, an active ketamine metabolite, in Wistar rats. *Pharmacol. Res. Perspect.* **3**, e00157 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank A. Keller for his assistance with the qEEG experiments, M. K. Lobo for assistance with the social defeat experiments, B. Alkondon for the rat hippocampal slice preparation and biocytin processing, V. Meadows and S. Krimmel for assistance with behavioural experiments and manuscript review, E. Pereira for critical comments on the manuscript and figures, D. Luckenbaugh for assistance with statistical analysis, and C. Moore for the small molecule X-ray crystallography. Research was supported by NIMH grants MH099345 and MH107615 to T.D.G. and MH086828 to S.M.T., and the NIA (R.M., I.W.W.), NIMH (C.A.Z.), and NCATS (C.J.T.) NIH intramural research programs. Receptor binding profiles and K_i determinations were supported by the NIMH Psychoactive Drug Screening Program, Contract HHSN-271-2008-025C, to B. L. Roth in conjunction with J. Driscoll. Initial synthesis of the ketamine metabolites used in this study was supported by NIA Contract HHSN2712010000081 to I.W.W.

Author Contributions P.Z., R.M., P.J.M., I.W.W., C.J.T., C.A.Z. and T.D.G. were responsible for the overall experimental design. P.J.M., Y.F. and C.J.T. synthesised the ketamine metabolites and deuterated ketamine derivatives, and provided mass spectrometer confirmations. Bioanalytical quantitation of ketamine and metabolites were performed by R.M., N.S.S. and K.S.S.D.. P.Z., P.G. and H.J.P. conducted and analysed the results of the behavioural and qEEG experiments. X.-P.H. supervised and analysed the results of the binding experiments. P.Y. performed the western blot experiments. E.X.A., M.A., J.F. and S.M.T. helped design and analyse the electrophysiology experiments, which were conducted by M.A., J.F. and S.M.T. G.I.E. and C.L.M. conducted and analysed the results of the i.v. self-administration. P.Z. and T.D.G. outlined and wrote the paper, which was reviewed by all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.D.G. (gouldlab@me.com).

METHODS

Animals. Mice and rats were acclimated to University of Maryland, Baltimore, vivarium for at least 7 days before experiments. Food and water were available *ad libitum*. CD-1 mice (males unless otherwise noted; 8–10 weeks old; Charles River Laboratories) were housed in groups of 4–5 per cage with a constant 12-h light/dark cycle (lights on/off at 07:00/19:00). For the social defeat experiments, 8–9-week-old male C57BL/6J mice (University of Maryland, Baltimore, veterinary resources breeding colony) and retired male CD-1 breeders (Charles River Laboratories) were used. For the whole-cell and field potential electrophysiological recordings, male Sprague–Dawley rats (postnatal day 24–35; Charles River) were used. All experimental procedures were approved by the University of Maryland, Baltimore Animal Care and Use Committee and were conducted in full accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals.

Drugs. (R,S)-ketamine, (S)-ketamine, desipramine, MK-801, PCP (Sigma-Aldrich), (R)-ketamine (Cayman Chemicals) and NBQX (National Institute of Mental Health Chemical Synthesis and Drug Supply Program) were dissolved in 0.9% saline. (2S,6S)-HNK, (2R,6R)-HNK and 6,6-dideuteroketamine hydrochloride were synthesized and characterized both internally at the National Center for Advancing Translational Sciences and at SRI International as described in Supplementary Information. Absolute and relative stereochemistry for (2S,6S)-HNK and (2R,6R)-HNK were confirmed by small molecule X-ray crystallography, as described in the Supplementary Information.

All drugs were dissolved in 0.9% saline, and administered intraperitoneally (i.p.) in a volume of 7.5 ml kg^{-1} of body mass by a male experimenter for the behavioural studies. Corticosterone (4-pregnen-11 β , 21-diol-3, 20-dione 21-hemisuccinate; Steraloids) was dissolved in tap water. For the electrophysiology recordings, test drugs were diluted in ACSF.

FST. Mice were tested in the FST 1 h and/or 24 h after injection. During the test, mice were subjected to a 6-min swim session in clear Plexiglass cylinders (30 cm height \times 20 cm diameter) filled with 15 cm of water ($23 \pm 1^\circ\text{C}$). The test was performed in normal light conditions (800 Lx). Sessions were recorded using a digital video camera. Immobility time, defined as passive floating with no additional activity other than that necessary to keep the animal's head above the water, was scored for the last 4 min of the 6-min test by a trained observer.

We conducted three different FST experiments where we used the AMPAR antagonist NBQX. In the first two experiments, we administered NBQX 10 min before ketamine, (2R,6R)-hydroxynorketamine, or vehicle and then tested the mice 1 h and 24 h later to assess whether AMPAR activity is necessary for the acute actions of these drugs, leading to both acute and sustained antidepressant effects. In the third experiment we first administered ketamine, (2R,6R)-HNK, or vehicle and 23.5 h later mice received either NBQX or vehicle. Thirty minutes later we tested these mice in the FST, to assess effects of AMPAR activity on sustained antidepressant actions.

Open-field test. Mice were placed into individual open-field arenas (50 \times 50 \times 38 cm (length \times width \times height); San Diego Instruments) for a 60-min habituation period. Mice then received an injection of the respective drug and assessed for locomotor activity for another 60 min. Distance travelled was analysed using TopScan v2.0 (CleverSys, Inc.).

NSF test. Mice were singly housed and food-deprived for 24 h in freshly made home-cages. Two normal chow diet pellets were placed on an inverted weighing-boat platform (10 \times 10 \times 1.5 cm) in the centre of an open-field arena (40 \times 40 cm). Thirty or sixty minutes (see figure legends) after drug administration, mice were introduced into a corner of the arena. The time needed for the mice to take a bite of food was recorded over a 10-min period by a trained observer. After the test, the mice were returned to their home cage containing pre-weighed food pellets, and latency to start biting the pellet, as well as consumption was recorded for a period of 10 min. There was no significant change in home cage latency or consumption in any of our experiments (data not shown).

Learned helplessness test. The learned helplessness model consisted of three different phases: inescapable shock training, learned helplessness screening, and the test. For the inescapable shock portion of the test (day 1), the animals were placed in one side of two-chambered shuttle boxes (34 \times 37 \times 18 cm (height \times width \times depth); Coulbourn Instruments), with the door between the chambers closed. After a 5-min adaptation period, 120 inescapable foot-shocks (0.45 mA, 15 s duration, randomized average inter-shock interval of 45 s) were delivered through the floor. During the screening session (day 2), the mice were placed in one of the two chambers of the apparatus for 5 min. A shock (0.45 mA) was then delivered, and the door between the two chambers was raised simultaneously. Crossing over into the second chamber terminated the shock. If the animal did not cross over, the shock terminated after 3 s. A total of 30 screening trials of escapable shocks were presented to each mouse with an average of 30-s delay

between each trial. Mice that developed helplessness behaviour (>5 escape failures during the last 10 screening shocks) received the assigned drug 24 h after screening (day 3). During the learned helplessness test phase (day 4), the animals were placed in the shuttle boxes and, after a 5-min adaptation period, a 0.45-mA shock was delivered concomitantly with door opening for the first five trials, followed by a 2-s delay for the next 40 trials. Crossing over to the second chamber terminated the shock. If the animal did not cross over to the other chamber, the shock was terminated after 24 s. A total of 45 trials of escapable shocks were presented to each mouse with 30-s inter-trial intervals. The number of escape failures was recorded for each mouse by automated computer software (Graphic State v3.1; Coulbourn Instruments).

Chronic social defeat stress and social interaction. The timeline for the social defeat experiments is presented in Extended Data Fig. 3a. Male C57BL/6J mice underwent a 10-day chronic social defeat stress model, as described elsewhere³⁴, with some modifications. In brief, experimental mice were introduced to the home cage (43 \times 11 \times 20 cm (length \times width \times height)) of a resident aggressive retired CD-1 breeder, pre-screened for aggressive behaviours, for 10 min. After this physical attack phase, mice were transferred and housed in the opposite side of the resident's cage divided by a Plexiglas perforated divider, to maintain continuous sensory contact. This process was repeated for 10 days. Experimental mice were introduced to a novel aggressive CD-1 mouse each day. On day 11, test mice were screened for susceptibility in a social interaction/avoidance choice test. The social interaction apparatus consisted of a rectangular three-chambered box (mouse conditioned-place preference chamber; Stoelting Co., see Extended Data Fig. 3b) containing two equal sized end-chambers and a smaller central chamber. The social interaction/avoidance choice test consisted of two 5-min phases. During the habituation phase, mice explored the empty apparatus. During the test phase, two small wire cages (Galaxy Cup, Spectrum Diversified Designs, Inc.), one containing a 'stranger' CD-1 mouse and the other one empty, were placed in the far corners of each chamber. The time spent interacting (nose within close proximity of the cage) with the stranger mouse versus the empty cage was analysed using TopScan video tracking software (CleverSys). Locomotor activity (total distance moved over 5 min) and number of total crosses into and out of the central chamber were also measured. The social interaction ratio was calculated by dividing the time spent interacting with the stranger by the time spent with the empty cage. Mice having a social interaction ratio higher than 1.0 were considered resilient, and mice with a social interaction ratio lower than 1.0 were considered susceptible. On day 13 resilient and susceptible mice received an i.p. injection of saline, (R,S)-KET (20 mg kg^{-1} ; chosen based on dose previously effective in C57BL/6J mice³⁴), MK-801 (0.1 mg kg^{-1}) or (2R,6R)-HNK (20 mg kg^{-1}). Mice were re-tested for social interaction/avoidance on day 15 (24 h after treatment).

Chronic corticosterone-induced anhedonia. Sucrose preference test. For assessing the baseline sucrose preference, mice were singly housed for 24 h and presented with two identical bottles containing either tap water or 1% sucrose solution. After baseline sucrose measurement, mice were re-group housed (5 mice per cage) and treated for 4 weeks with corticosterone (25 $\mu\text{g ml}^{-1}$ equivalent) given in water bottles. Before initiation of any behavioural measurements, animals were weaned off corticosterone treatment; 3 days corticosterone 12.5 $\mu\text{g ml}^{-1}$ and 3 days corticosterone 6.25 $\mu\text{g ml}^{-1}$, followed by 1 week of complete withdrawal. Mice were subsequently singly housed in freshly made home cages and provided with two bottles containing either tap water or 1% sucrose solution. Twenty-four hours later, mice that developed the anhedonia phenotype ($<70\%$ sucrose preference) were treated with saline or (2R,6R)-HNK (10 mg kg^{-1}) and sucrose preference was measured after an additional 24 h.

Female urine sniffing test. A separate cohort of mice was treated with the same chronic corticosterone administration model as described above but assessed for female urine sniffing preference as a measure of hedonic behaviour³⁵. Mice were singly housed in freshly made home cages for a habituation period of 10 min. Subsequently, one plain cotton tip was secured on the centre of the cage wall and mice were allowed to sniff and habituate to the tip for a period of 30 min. Then, the plain cotton tip was removed and replaced by two cotton tip applicators, one infused with fresh female mouse oestrus urine and the other with fresh male mouse urine. These applicators were presented and secured at the two corners of the cage wall simultaneously. Sniffing time for both female and male urine was scored by a trained observer for a period of 3 min. Twenty-four hours later, mice that developed the anhedonia phenotype ($<65\%$ female urine preference; susceptible phenotype), as well as mice that did not develop the anhedonia phenotype ($>75\%$ female urine preference; resilient phenotype) were treated with either saline or (2R,6R)-HNK (10 mg kg^{-1}). Mice were re-tested for female urine preference 24 h later.

Pre-pulse inhibition. Mice were individually tested in acoustic startle boxes (SR-LAB, San Diego Instruments). After drug administration, mice were placed in the startle chamber for a 30-min habituation period. The experiment started

with a further 5-min adaptation period during which the mice were exposed to a constant background noise (67 dB), followed by five initial startle stimuli (120 dB, 40 ms duration each). Subsequently, animals were exposed to four different trial types: pulse alone trials (120 dB, 40 ms duration), three pre-pulse trials of 76 and 81 of white noise bursts (20 ms duration) preceding a 120 dB pulse by 100 ms, and background (67 dB) no-stimuli trials. Each of these trials was randomly presented five times. The dose of ketamine (30 mg kg⁻¹) was selected based on a dose-response experiment we performed in a previous study³⁶. The percentage pre-pulse inhibition was calculated using the following formula: ((magnitude on pulse-alone trial – magnitude on pre-pulse + pulse trial)/magnitude on pulse-alone trial) × 100.

Drug discrimination. Mice were food-restricted until they reached 85% of their initial body weight and were maintained at 85% throughout the duration of the experiment. Animals were trained to lever press for food (20 mg sucrose pellets; TestDiet) in standard two-lever operant conditioning chambers (Coulbourn Instruments), under a fixed-ratio 5 schedule of reinforcement (FR5) in daily 30-min sessions. After stable responding was maintained over three consecutive sessions, mice were trained to discriminate ketamine (10 mg kg⁻¹) from saline under a double alternation schedule (that is, ketamine, ketamine, saline, saline), which required on average 40 training sessions. Mice received either ketamine (10 mg kg⁻¹; i.p.) or saline 15 min before the start of the 30-min session. Responding to the correct lever resulted in the delivery of a reward, while incorrect responding reset the FR for correct lever-responding. Drug discrimination test sessions were conducted when mice reached the following criteria: (1) first FR5 completed on the correct lever, and (2) ≥85% correct lever responding over the entire session. Fifteen minutes prior to the 30-min test sessions mice received either saline, ketamine (10 mg kg⁻¹), PCP (3 mg kg⁻¹) or (2R,6R)-HNK (10 and 50 mg kg⁻¹). At this stage, completion of the FR5 schedule on either lever resulted in the delivery of food reward. Lever response and pellet delivery were monitored and controlled by an automated computer system (Graphic State v3.1; Coulbourn Instruments).

Intravenous drug self-administration. *Apparatus.* Each operant chamber was equipped with one lever, a dipper liquid delivery system, a 22-gauge liquid swivel and a syringe pump (located outside the chamber). The lever was a balanced rocker arm that broke an infrared photo beam when 0.5 g of force was applied. Two stimulus lights were used; one was positioned to illuminate the translucent lever and the other was positioned above the liquid delivery recess. The lever light was illuminated during periods of water or drug availability; the second light was illuminated during water or drug delivery. The system was controlled by an integrated Coulbourn environmental control system and Med Associates interface.

Water training. Mice were first trained to complete an operant response for water reinforcement. Completion of the response requirements on the lever illuminated stimulus lights above a spout and delivered a small amount of water. Initially, the response requirement was one lever press (FR1); after completion of each 50 reinforcements the fixed ratio requirement was increased by one (FRX + 1). Mice were trained for 24 h per day for 4 days with free access to food.

Surgery. After completion of the water training, mice were surgically prepared with a catheter implanted in the jugular vein. Surgical procedures were performed under ketamine- (90 mg kg⁻¹, i.p.) and xylazine- (16 mg kg⁻¹, i.p.) induced anaesthesia. Silastic tubing (0.012 inch (0.30 mm) inner diameter) was implanted in the right jugular vein to the level of the atrium, passed subcutaneously and exited in the midscapular region. The catheter was connected to a tether/swivel system that was mounted to the skull of the mouse with dental cement.

Intravenous drug self-administration. Seven days after surgery, mice were placed in the operant chamber and given access (FR4) to 0.32, 1.0, 3.2 or 0 (saline) mg kg⁻¹ drug per infusion for 5 days at each dose. Completion of each FR resulted in the illumination of the overhead house light and the stimulus light above the spout. Infusions of 5–8 µl (based on body weight) were given over a period of 15 s. A 30-s time-out period, during which house and stimulus lights were out, followed the completion of each infusion. Each mouse had access to drug for 6 h per day and free access to food and water 24 h per day. A 12-h light/dark cycle was maintained (lights on/off at 07:00/19:00). Each animal remained in its operant chamber for the duration of the experiment. A stimulus light illuminating the lever signalled drug availability. Only those animals with patent catheters at the end of the experiment were included in the analysis. The average number of reinforcements and drug intake during the last 3 days at each dose were used as dependent measures.

Rotarod. The rotarod test was conducted to compare the effects of ketamine, (2S,6S)-HNK and (2R,6R)-HNK on motor coordination. The experiment consisted of two phases: training phase (4 days) and a test phase (1 day). On each of the training days five trials (trial time: 3 min) were conducted with an inter-trial interval of two min. Mice were individually placed on the rotarod apparatus (IITC Life Science) and the rotor (3.75 inch diameter) accelerated from 5 to 20 r.p.m. over

a period of three minutes. Latency to fall was recorded for each trial. Animals with an average of <100 s of latency to fall during the last training day were excluded from the experiment. On the test day (day 5), mice received (i.p.) injections of saline, (R,S)-KET (10 mg kg⁻¹), (2S,6S)-HNK (25 or 125 mg kg⁻¹) or (2R,6R)-HNK (25 or 125 mg kg⁻¹) and were tested in the rotating rod 5, 10, 15, 20, 30 and 60 min after injection using the same procedure described for the training days.

Tissue distribution and clearance measurements of ketamine and metabolites. At 10, 30, 60, or 240 min after drug administration mice were exposed to 3% isoflurane and subsequently decapitated. Trunk blood was collected in EDTA-containing tubes and centrifuged at 5,938g for 6 min (4 °C). Plasma was collected and stored at –80 °C until analysis. Whole brains were simultaneously collected, rinsed with PBS, immediately frozen in dry ice and stored at –80 °C until analysis.

The concentrations of (R,S)-ketamine and 6,6-dideuteroketamine and their respective metabolites in plasma and brain tissue were determined by achiral liquid chromatography-tandem mass spectrometry following a previously described method³³, with slight modifications. The analysis was accomplished using an Eclipse XDB-C18 guard column (4.6 × 12.5 mm) and a Varian Pursuit XRs C18 analytical column (250 × 4.0 mm ID, 5 µm; Varian). The mobile phase consisted of ammonium acetate (5 mM, pH 7.6) as component A and acetonitrile as component B. A linear gradient was run as follows: 0 min 20% B; 5 min 20% B; 15 min 80% B; 20 min 20% B at a flow rate of 0.4 ml min⁻¹. The total run time was 30 min per sample. For plasma and brain samples, the calibration standards ranged from 10,000 ng ml⁻¹ to 19.5 ng ml⁻¹ for (R,S)-ketamine, (R,S)-norketamine, (2R,6R;2S,6S)-HNK, (R,S)-dehydronorketamine, (R,S)-d₂-ketamine, (R,S)-d₂-norketamine and d-(2S,6S;2R,6R)-HNK. The quantification of (R,S)-ketamine, (R,S)-d₂-ketamine and their respective metabolites was accomplished by calculating area ratios using d₄-ketamine (10 µl of 10 µg ml⁻¹ solution) as the internal standard. The MS/MS analysis was performed using a triple quadrupole mass spectrometer model API 4000 system from Applied Biosystems/MDS Sciex equipped with Turbo Ion Spray (TIS) (Applied Biosystems). The data was acquired and analysed using Analyst version 1.4.2 (Applied Biosystems). Positive electrospray ionization data were acquired using multiple reaction monitoring (MRM) using the following transitions for (R,S)-ketamine studies: 238 → 125 (ketamine); 224 → 125 (norketamine); 222 → 177 (dehydronorketamine); 240 → 125 (HNK); 254 → 151 (HK) and (R,S)-d₂-ketamine studies: 240 → 125 (d₂-ketamine); 226 → 125 (d₂-norketamine); 223 → 178 (d-dehydronorketamine); 241 → 125 (d-(2,6)-HNK); 242-125 (d₂-(2,5)-HNK); (d₂-(2,4)-HNK) and 255 → 151 (d-(2,6)-HK).

MK-801 displacement binding. NMDAR binding assays were performed according to ref. 37, with minor modifications. Rat brains were homogenized and membrane fractions were collected. Aliquoted membranes were stored at –80 °C until use. Membrane pellets were washed five times with an ice-cold buffer (20 mM HEPES, 1 mM EDTA, pH 7.0) before use. The binding assays were set up in 96-well plates using 5 nM [³H]MK-801 and rat brain membranes (100 µg per well) in a final volume of 125 µl per well in the NMDAR binding buffer (20 mM HEPES, 1 mM EDTA, 100 µM glutamate, 100 µM glycine, pH 7.0). Test compounds were first distributed in 96-well plates (25 µl per well at 5 × final concentrations ranging from 0.1 nM to 10 µM, 11 points) in triplicate. The radioligand, [³H]MK-801, was added (50 µl per well at 2.5 × of final concentration of 5 nM) to all wells. Reactions started with the addition of 50 µl rat brain membrane and were incubated for 1 h in the dark at room temperature. The reactions were harvested via rapid filtration onto Whatman GF/B glass fibre filters pre-soaked with 0.3% polyethyleneimine using a 96-well Brandel harvester, followed by three quick washes each with 500 µl chilled wash buffer (50 mM Tris HCl, pH 7.4). Filters were microwave-dried and scintillation cocktail was then melted onto the filter mates on a hot plate. The radioactivity retained on the filters was counted in a MicroBeta scintillation counter. All assays were performed in triplicates.

Western blots. To purify synaptoneurosomes, mouse prefrontal cortex and hippocampi were dissected and homogenized in Syn-PER Reagent (ThermoFisher Scientific; 87793) with 1 × protease and phosphatase inhibitor cocktail (ThermoFisher Scientific; 78440). The homogenate was centrifuged for 10 min at 1,200g at 4 °C. The supernatant was centrifuged at 15,000g for 20 min at 4 °C. After centrifugation, the pellet (synaptosomal fraction) was re-suspended and sonicated in N-PER Neuronal Protein Extraction Reagent (ThermoFisher Scientific; 87792). Protein concentration was determined via the BCA protein assay kit (ThermoFisher Scientific; 23227). Equal amount of proteins (10–40 µg as optimal for each antibody) for each sample was loaded into NuPage 4–12% Bis-Tris gel for electrophoresis. Gel transfer was performed with the TransBlot Turbo Transfer System (Bio-Rad) Nitrocellulose membranes with transferred proteins were blocked with 5% milk in TBST (TBS plus 0.1% Tween-20) for 1 h and kept with primary antibodies overnight at 4 °C. The following primary antibodies were used: phospho-eEF2 (at Thr56; Cell Signaling Technology; 2331), total eEF2 (Cell Signaling Technology; 2332), phospho-mTOR (at Ser2448; Cell Signaling Technology; 2971), total mTOR

(Cell Signaling Technology; 2983), GluA1 (Cell Signaling Technology; 2983), GluA2 (Cell Signaling Technology; 13607), BDNF (Santa Cruz Biotechnology; sc-546), and GAPDH (Abcam; ab8245). The next day, blots were washed three times in TBST and incubated with horseradish peroxidase conjugated anti-mouse or anti-rabbit secondary antibody (1:5,000 to 1:10,000) for 1 h. After three final washes with TBST, bands were detected using enhanced chemiluminescence (ECL) with the Syngene Imaging System (G:Box ChemiXX9). After imaging, the blots were incubated in the stripping buffer (ThermoFisher Scientific; 46430) for 10–15 min at room temperature followed by three washes with TBST. The stripped blots were incubated in blocking solution for 1 h and incubated with the primary antibody directed against total levels of the respective protein or GAPDH for loading control. Densitometric analysis of phospho- and total immunoreactive bands for each protein was conducted using Syngene's GenTools software. The values for the phosphorylated forms of proteins were normalized to phosphorylation-independent levels of the same protein. Phosphorylation-independent levels of proteins were normalized to GAPDH. Fold change was calculated by normalization to saline-treated control group for each protein or phosphoprotein.

qEEG. Surgery. qEEG experiments were performed according to ref. 38, with minor modifications. Mice were anaesthetized with isoflurane (3.5%) and maintained under anaesthesia (2–2.5%) throughout the surgery. Mice received analgesia (carprofen, 5 mg kg⁻¹, i.p.) before the start of surgery. An F20-EET radio-telemetric transmitter (Data Sciences International) was placed subcutaneously and its leads implanted over the dura above the frontal cortex (1.7 mm anterior to bregma) and the cerebellum (6.4 mm posterior to bregma). Animals recovered from surgery for 7 days before recordings.

qEEG recordings. For comparisons between saline, ketamine, and (2R,6R)-HNK, mice were singly housed and acclimated to the behavioural room for 24 h before qEEG recordings. qEEGs were recorded using the Dataquest A.R.T. acquisition system (Data Sciences International) with frontal qEEG recordings referenced to the cerebellum. Baseline qEEG (30 min) recordings were followed by an i.p. injection of saline, ketamine (10 mg kg⁻¹) or (2R,6R)-HNK (10 mg kg⁻¹) and a further 60 min of post-injection recordings. To assess effects of NBQX on (2R,6R)-HNK-induced changes in qEEG oscillations, mice were acclimated to the behavioural room 1.5–2 h before recordings. Baseline (30 min) recordings were followed by an i.p. injection of either saline or NBQX (10 mg kg⁻¹) and 30 min later mice received an injection (i.p.) of (2R,6R)-HNK (10 mg kg⁻¹) and recordings continued for 60 min after injection. **In vivo data analysis.** qEEGs were analysed using custom-written MATLAB scripts (Version 2012a, Mathworks) and the mtspecgram routine in the Chronux Toolbox (<http://chronux.org>³⁹). Oscillation power in each bandwidth (delta = 1–3 Hz; theta = 4–7 Hz; alpha = 8–12 Hz; beta = 13–29 Hz; gamma = 30–80 Hz) was computed in 10-min bins from spectrograms for each animal.

Field recordings. Removal of the rat brains, as well as dissection of the hippocampi were performed in ice-cold ACSF bubbled with 95% O₂, 5% CO₂. The ACSF contained (in mM): 124 NaCl, 3 KCl, 1.25 NaH₂PO₄, 1.5 MgCl₂, 2.5 CaCl₂, 26 NaHCO₃ and 10 glucose. Hippocampal slices were cut at 400 µm using a vibratome and kept in a holding chamber at the interface of ACSF and humidified 95% O₂, 5% CO₂ for at least 1 h. For the fEPSPs, slices were transferred to a submersion-type recording chamber and perfused with ACSF (1–2 ml min⁻¹; room temperature). Picrotoxin (0.1 mM), CGP52432 (2 µM) and APV (80 µM) were added to block GABA_A, GABA_B and NMDA receptors respectively. Concentric bipolar tungsten electrodes were placed in stratum radiatum to stimulate the Schaffer collateral afferents. Extracellular recording pipettes were filled with ACSF (3–5 MΩ) and placed in stratum radiatum of area CA1. Field potentials were evoked by monophasic stimulation (100 µs duration) at 0.1 Hz. The stimulus intensity was set at 150% of threshold intensity, resulting in fEPSPs amplitude of 0.1–0.3 mV. A stable baseline was recorded for at least 10 min. Vehicle and (2R,6R)-HNK were applied by perfusion over a period of 1 h followed by washout using ACSF. For AMPAR-mediated responses, peak fEPSP amplitudes and slopes, measured over a window of 1–4 ms following the rising phase of the response, are reported as percentage change from baseline. DNQX (50 µM) was bath-applied to ensure AMPA-mediated responses. Experiments were performed and analysed blind to treatment groups, using pCLAMP software (Molecular Devices).

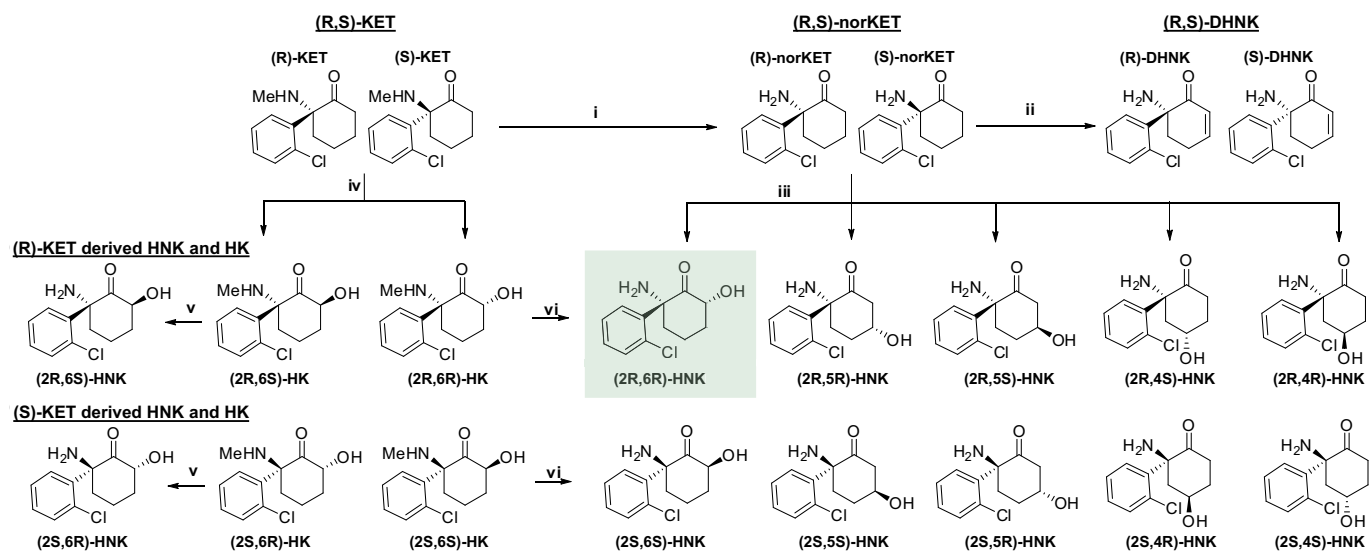
Whole-cell patch-clamp recordings. Rats were euthanized by CO₂ asphyxiation followed by decapitation. Removal of the brains, as well as dissection and slicing of the hippocampi were performed in an ice-cold solution consisting of a mixture of equal parts of regular ACSF and sucrose-containing ACSF. ACSF was composed of (in mM): 125 NaCl, 26 NaHCO₃, 2.5 KCl, 1.25 NaH₂PO₄, 2 CaCl₂,

1 MgCl₂, and 25 glucose. Sucrose-containing ACSF was composed of (in mM): 230 sucrose, 2.5 KCl, 1.25 NaH₂PO₄, 26 NaHCO₃, 0.5 CaCl₂, 10 MgSO₄, and 10 glucose. Hippocampal slices of 300-µm thickness were cut using a vibratome (Leica VT1000S; Leica Microsystems Inc.) and transferred to an immersion chamber containing regular ACSF that was continuously bubbled with 95% O₂, 5% CO₂ and maintained in a water bath at 30 °C.

At the time of recordings, hippocampal slices were transferred to a 1-ml recording chamber, where they were superfused at 2 ml min⁻¹ with ACSF that was continuously bubbled with 95% O₂, 5% CO₂. In all experiments, ACSF used to superfuse the slices contained the muscarinic antagonist atropine (0.5 µM) and the GABA_A receptor antagonist picrotoxin (50 µM). Whole-cell patch-clamp recordings were obtained from the soma of CA1 stratum radiatum interneurons in hippocampal slices according to standard patch-clamp techniques using an EPC9 amplifier (HEKA Elektronik). The signals were filtered at 3 kHz and analysed using pCLAMP 10.3 or WinEDR v3.2.6 (University of Strathclyde, UK). The patch-clamp pipettes were pulled from a borosilicate glass capillary (1.2-mm OD) and had resistances between 3 and 5 MΩ when filled with internal solution. The internal pipette solution contained (in mM): 10 ethylene-glycol-bis(3-amino-ethylether)-N,N'-tetraacetic acid, 10 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid, 130 Cs-methane sulfonate, 10 CsCl, 2 MgCl₂, 5 lidocaine N-ethyl bromide, and 0.5% biocytin (pH adjusted to 7.3 with 340 mOsm CsOH). A specially adapted U-tube developed in the Albuquerque laboratory was used to apply NMDA (50 µM) to the neurons for the NMDA-evoked current experiments. NMDA-evoked currents and spontaneous AMPA-mediated excitatory postsynaptic currents (sEPSCs) were recorded at -40 mV and -60 mV respectively. A single neuron was studied in each slice. All experiments were carried out at room temperature (20–22 °C). The peak amplitude of NMDA-evoked currents was analysed using the pCLAMP v10.3 software, with baseline determined as the mean value obtained before drug application and the final (16 min) washout time point. Frequency and peak amplitude of AMPA sEPSCs were analysed using WinEDR v3.2.6.

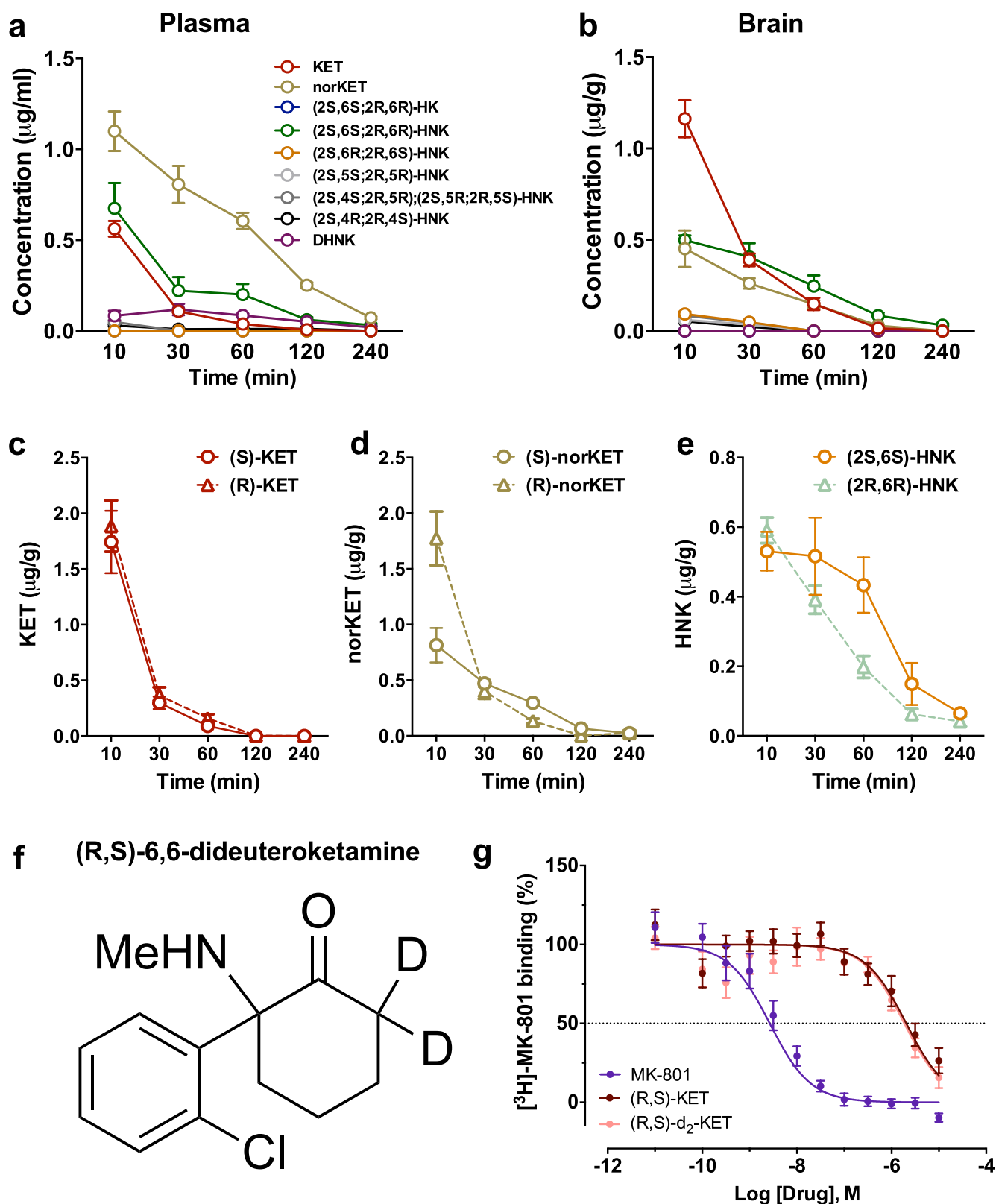
Statistical analyses. Required sample sizes were estimated based on our past experience performing similar experiments. Experimentation and analysis were performed in a manner blinded to treatment assignments in all experiments with the exception of the whole-cell patch-clamp recordings and intravenous drug self-administration. For all blinded experiments, mice were randomly assigned to treatment groups. Statistical analyses were performed using GraphPad Prism software v6. By pre-established criteria, values greater than ±2 s.d. from individual group means were excluded from the analyses. All statistical tests were two-tailed, and significance was assigned at $P < 0.05$. Normality and equal variances between group samples were assessed using the Kolmogorov–Smirnov and Brown–Forsythe tests respectively. When normality and equal variance between sample groups was achieved, ANOVAs were followed by a Holm–Šidák post-hoc comparison when significance was reached, and significant results are indicated with asterisks in the figures. As a secondary analysis, pairwise comparisons at each equivalent dose were performed followed by multiple comparison corrections, where appropriate, and are reported in Supplementary Information Table 1. Where normality or equal variance of samples failed, non-parametric one-way ANOVAs (Kruskal–Wallis one-way ANOVA on ranks or Friedman repeated-measures one-way ANOVA on ranks) were performed, followed by Dunn's correction. For assessment of the NSF test results, Kaplan–Meier survival analysis was used followed by the Mantel–Cox log-rank test. The sample sizes (biological replicates), specific statistical tests used, and the main effects of our statistical analyses for each experiment are reported in Supplementary Information Table 1.

34. Donahue, R. J., Muschamp, J. W., Russo, S. J., Nestler, E. J. & Carlezon, W. A. Jr. Effects of striatal ΔFosB overexpression and ketamine on social defeat stress-induced anhedonia in mice. *Biol. Psychiatry* **76**, 550–558 (2014).
35. Malkesman, O. *et al.* The female urine sniffing test: a novel approach for assessing reward-seeking behavior in rodents. *Biol. Psychiatry* **67**, 864–871 (2010).
36. Zanos, P. *et al.* The prodrug 4-chlorokynurenine causes ketamine-like antidepressant effects, but not side effects, by NMDA/glycine-site inhibition. *J. Pharmacol. Exp. Ther.* **355**, 76–85 (2015).
37. Chiu, J. *et al.* Chronic ethanol exposure alters MK-801 binding sites in the cerebral cortex of the near-term fetal guinea pig. *Alcohol* **17**, 215–221 (1999).
38. Raver, S. M., Haughwout, S. P. & Keller, A. Adolescent cannabinoid exposure permanently suppresses cortical oscillations in adult mice. *Neuropsychopharmacology* **38**, 2338–2347 (2013).
39. Bokil, H., Andrews, P., Kulkarni, J. E., Mehta, S. & Mitra, P. P. Chronux: a platform for analyzing neural signals. *J. Neurosci. Methods* **192**, 146–151 (2010).



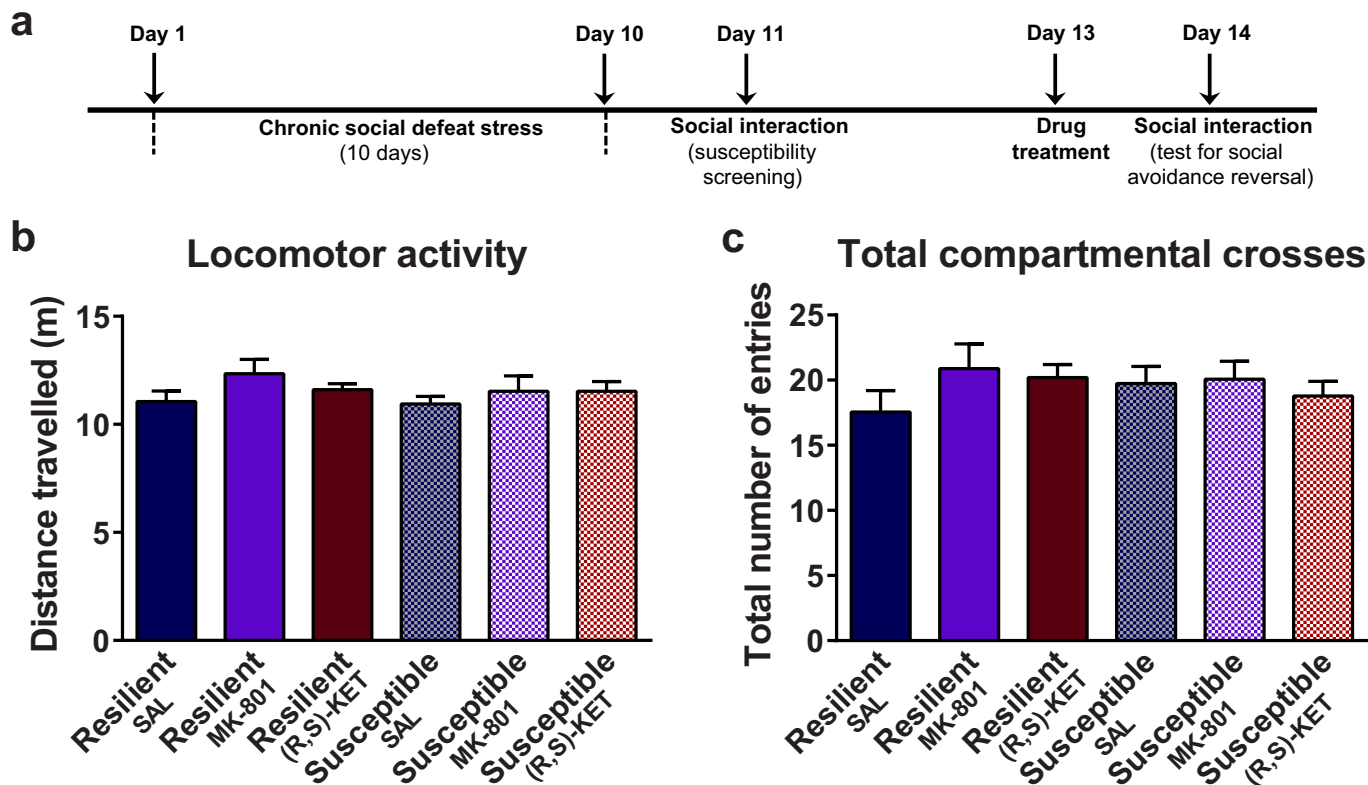
Extended Data Figure 1 | The metabolic transformations of ketamine *in vivo*. Ketamine is metabolised *in vivo* via P450 enzymatic transformations. **i**, (R,S)-KET is selectively demethylated to give (R,S)-norketamine (norKET). **ii**, NorKET can be then dehydrogenated to give (R,S)-dehydronorketamine (DHNK). **iii**, Alternatively, norKET can be hydroxylated to give the hydroxynorketamines (HNKs). **iv**, (R,S)-KET

can also be hydroxylated at the 6 position to give either the *E*-6-hydroxyketamine ((2S,6R;2R,6S)-HK) or *Z*-6-hydroxyketamine ((2S,6S;2R,6R)-HK). **v**, Demethylation of (2S,6R;2R,6S)-HK yields the production of (2S,6R;2R,6S)-HNK. **vi**, Demethylation of (2S,6S;2R,6R)-HNK further gives (2S,6S;2R,6R)-HNK.

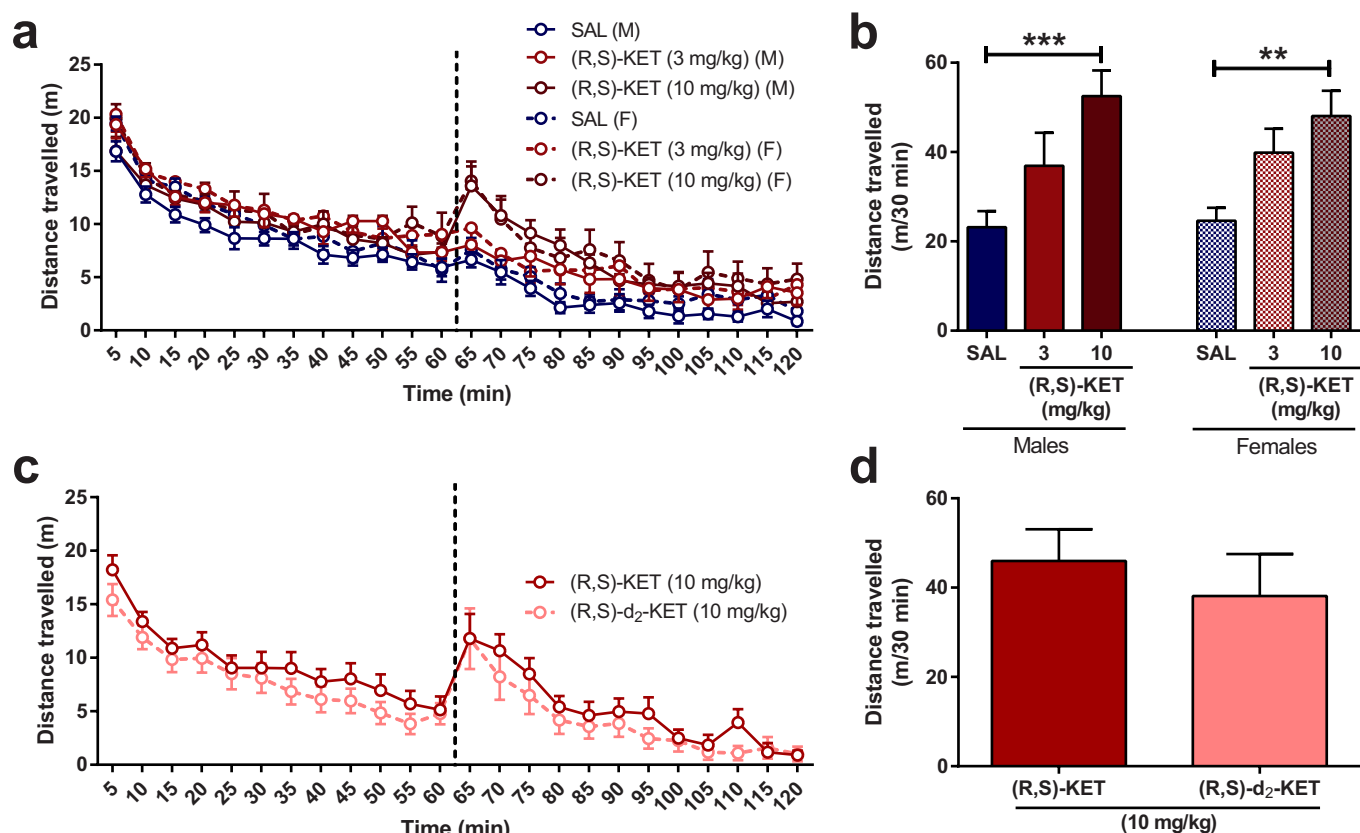


Extended Data Figure 2 | Circulating levels of ketamine and its metabolites following i.p. administration in mice. a, b, Plasma (a) and brain (b) levels of ketamine and its metabolites after administration of (R,S)-KET (10 mg kg^{-1}) in mice. **c, d,** Brain levels of KET (c), norKET (d) and HNK (e) following administration of (S)- and (R)-KET. **f, g,** Chemical

structure of (R,S)-6,6-dideuteroketamine ((R,S)-d₂-KET) (f), which displaces [^3H]MK-801 binding with a similar affinity to (R,S)-KET: $K_i = 799 \text{ nM}$; (R,S)-d₂-KET: $K_i = 883 \text{ nM}$) (g). See Supplementary Table 1 for statistical analyses and *n* numbers.

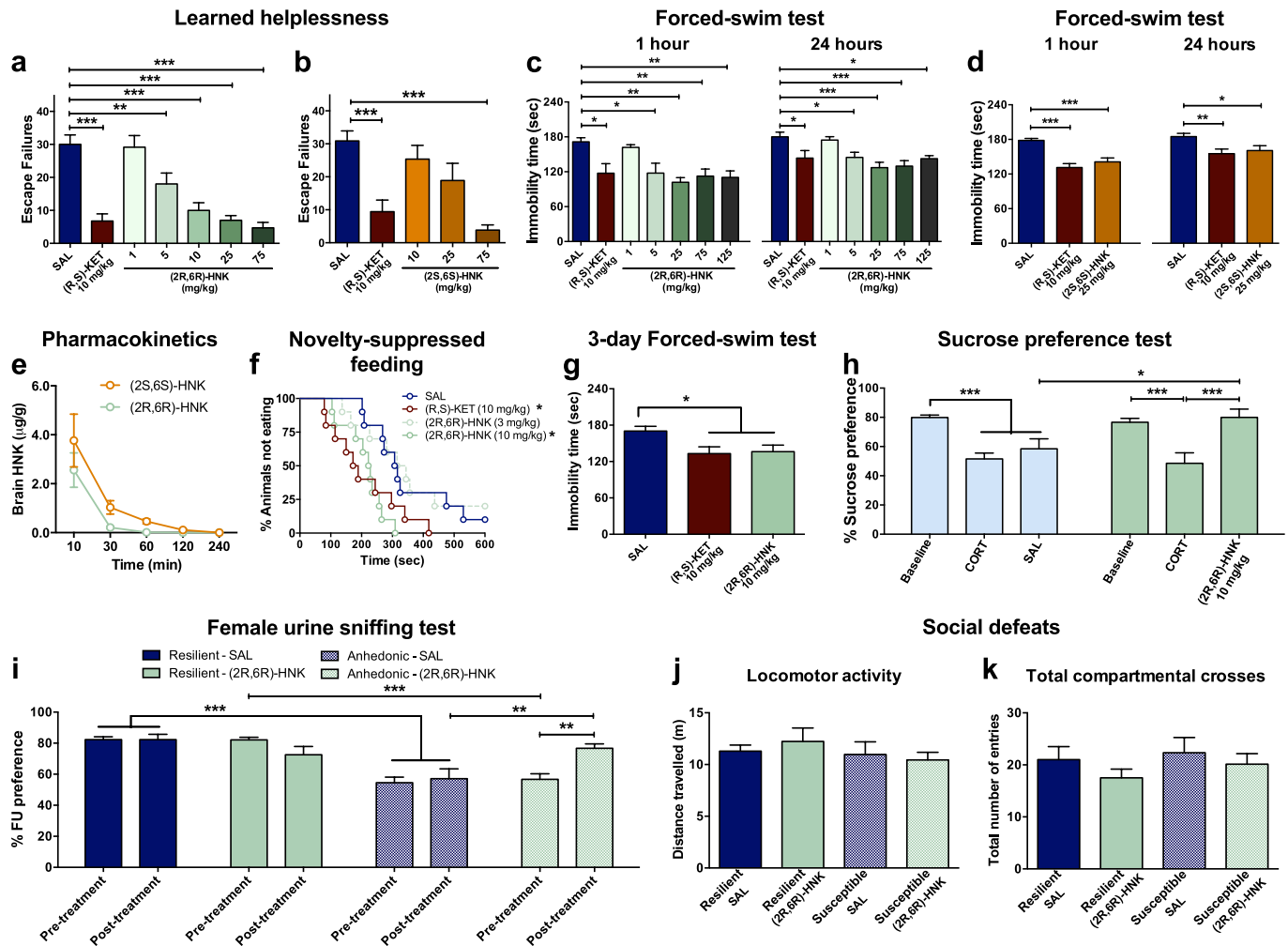


Extended Data Figure 3 | Additional social defeat stress data. **a**, Chronic social defeat stress and social interaction/avoidance test timeline. **b**, **c**, 24 h after administration, neither (R,S)-KET nor MK-801 affected locomotor activity (**b**) or total number of compartmental crosses in the social interaction apparatus (**c**). Data are mean \pm s.e.m. *** $P < 0.001$. See Supplementary Table 1 for statistical analyses and n numbers.



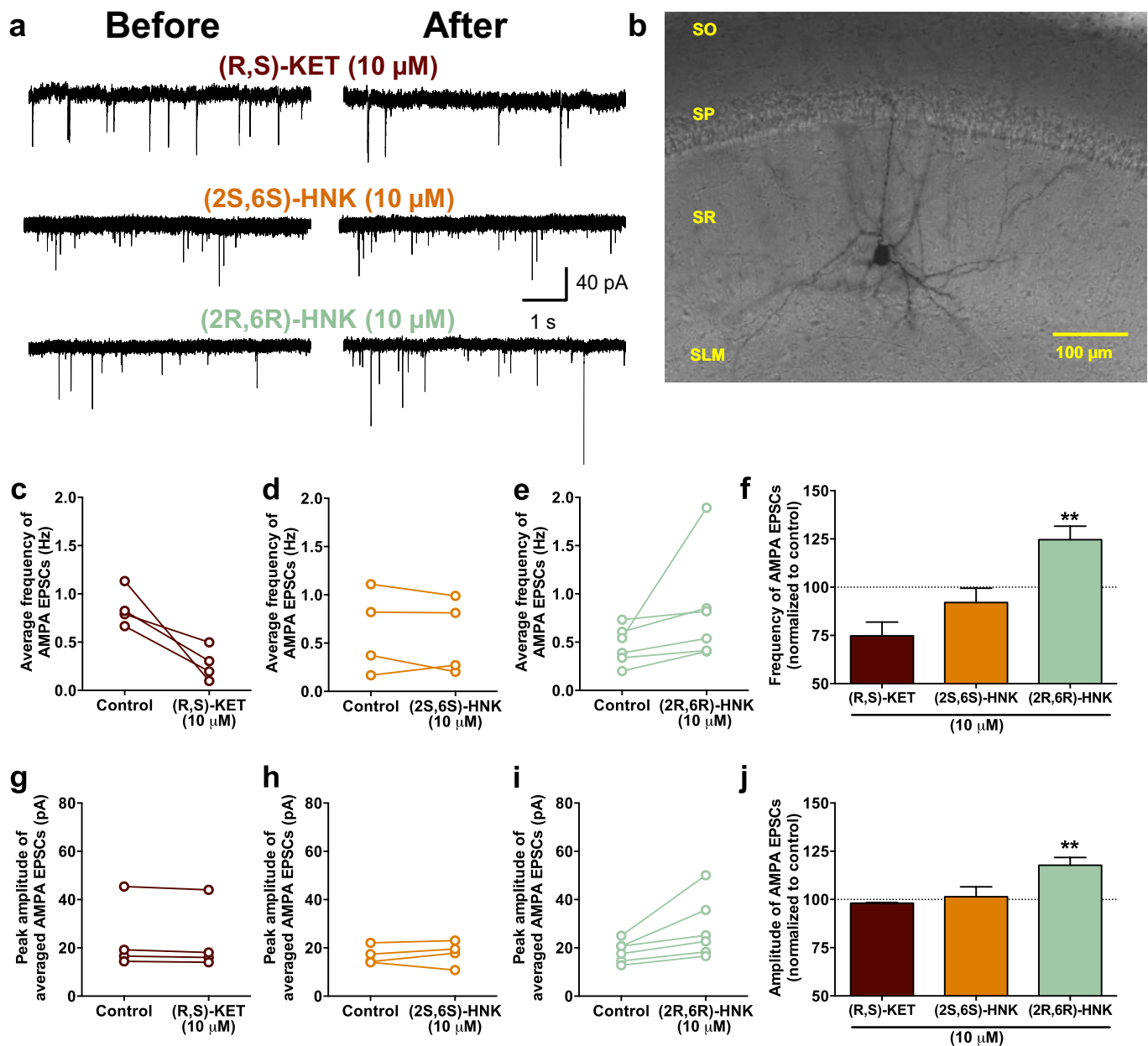
Extended Data Figure 4 | Locomotor effects of (R,S)-KET and (R,S)-d₂-KET. After recording baseline activity for 60 min, mice received drug (marked by a vertical dashed line) and locomotor activity was monitored for another 1 h. **a, b,** Administration of (R,S)-KET (10 mg kg⁻¹), induced hyperlocomotor responses equally in both male and female mice.

c, d, (R,S)-KET and (R,S)-d₂-KET were equally potent in inducing a hyperlocomotor response at the dose of 10 mg kg⁻¹. Data are mean ± s.e.m. **P* < 0.05, ***P* < 0.01 (see Supplementary Table 1 for statistical analyses and *n* numbers).



Extended Data Figure 5 | Acute and sustained antidepressant and anti-anhedonic effects of (2R,6R)- and (2S,6S)-HNK. **a**, A single injection of (2R,6R)-HNK resulted in dose-dependent antidepressant-like responses in the learned helplessness test at the doses of 5–75 mg kg⁻¹. **b**, A single injection of (2S,6S)-HNK induced antidepressant-like effects in the learned helplessness test at the dose of 75 mg kg⁻¹. **c**, Administration of (2R,6R)-HNK induced dose-dependent antidepressant effects in the 1- and 24-h FST. **d**, Administration of (2S,6S)-HNK at the dose of 25 mg kg⁻¹ induced antidepressant effects in the 1- and 24-h FST. **e**, Despite the greater antidepressant efficacy of (2R,6R)-HNK, administration of (2S,6S)-HNK (HNK) results in higher brain hydroxynorketamine levels compared to (2R,6R)-HNK. **f**, (2R,6R)-HNK manifested dose-dependent

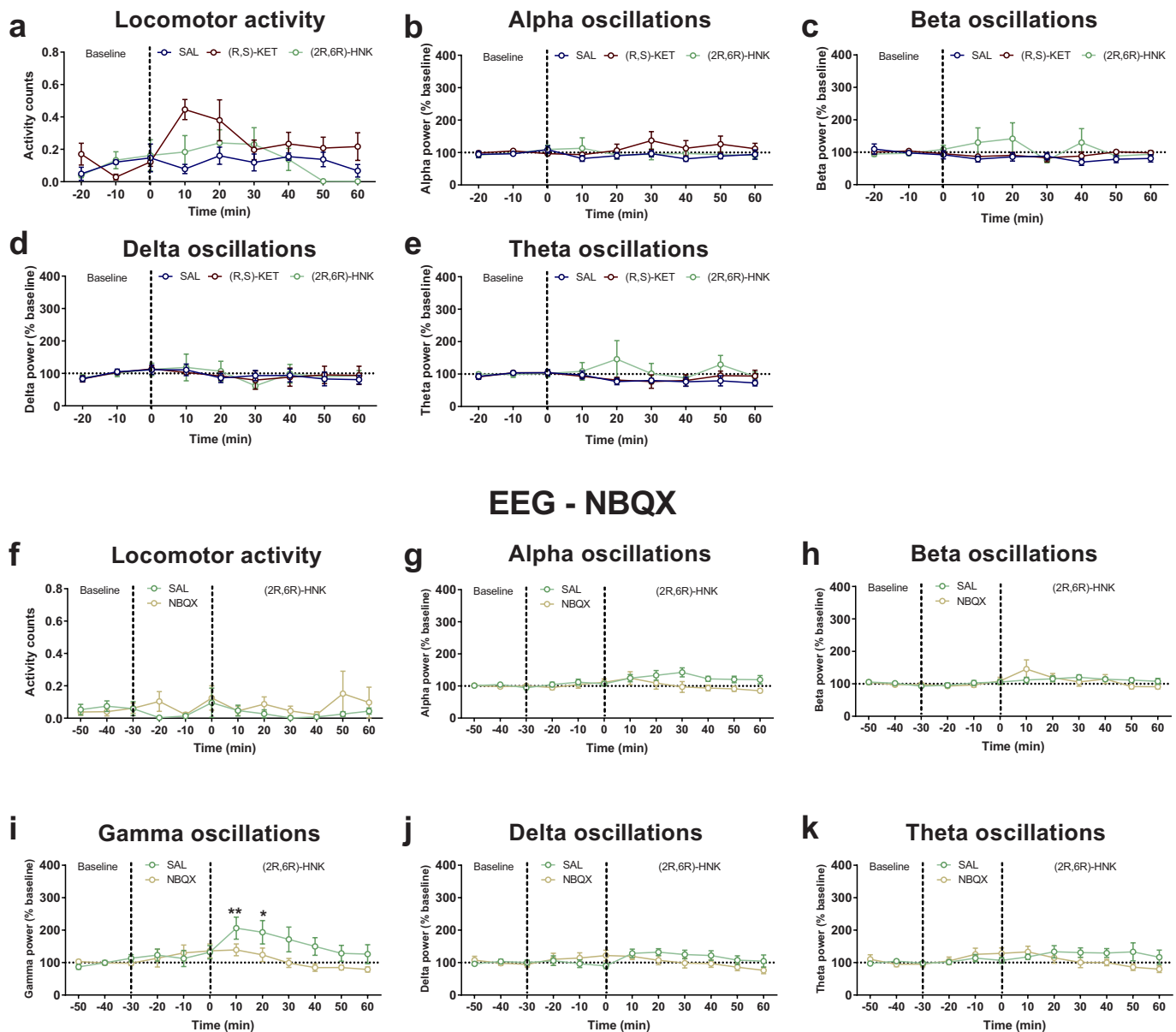
antidepressant-like effects in the NSF test. **g**, Similar to (R,S)-KET, the antidepressant-like effects of (2R,6R)-HNK in the FST persisted for at least 3 days after treatment. **h**, A single administration of (2R,6R)-HNK reversed chronic corticosterone-induced decreases in sucrose preference. **i**, A single administration of (2R,6R)-HNK reversed chronic corticosterone-induced decrease in female urine sniffing preference, specifically in mice that developed an anhedonic phenotype. Administration of (2R,6R)-HNK was not associated with changes in locomotor activity (**j**) or total compartmental crosses in the social interaction test after chronic social defeat stress (**k**). Data are mean \pm s.e.m. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ (see Supplementary Table 1 for statistical analyses and n numbers).



Extended Data Figure 6 | (2R,6R)-HNK rapidly increases the frequency and amplitude of AMPAR spontaneous excitatory postsynaptic currents in the hippocampus. **a**, Representative traces of spontaneous excitatory postsynaptic currents (sEPSCs) mediated via AMPARs during baseline (before) and 20 min after drug administration. **b**, Example CA1 stratum radiatum interneuron recorded from a rat hippocampal slice.

c–j, Twenty-minute exposure of (2R,6R)-HNK (**e**, **i**), but not (R,S)-KET (**c**, **g**) or (2S,6S)-HNK (**d**, **h**), increased AMPA sEPSCs frequency and amplitude compared to baseline. Data are mean \pm s.e.m. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ (see Supplementary Table 1 for statistical analyses and n numbers). SLM, stratum lacunosum-moleculare; SO, stratum oriens; SP, stratum pyramidale; SR, stratum radiatum.

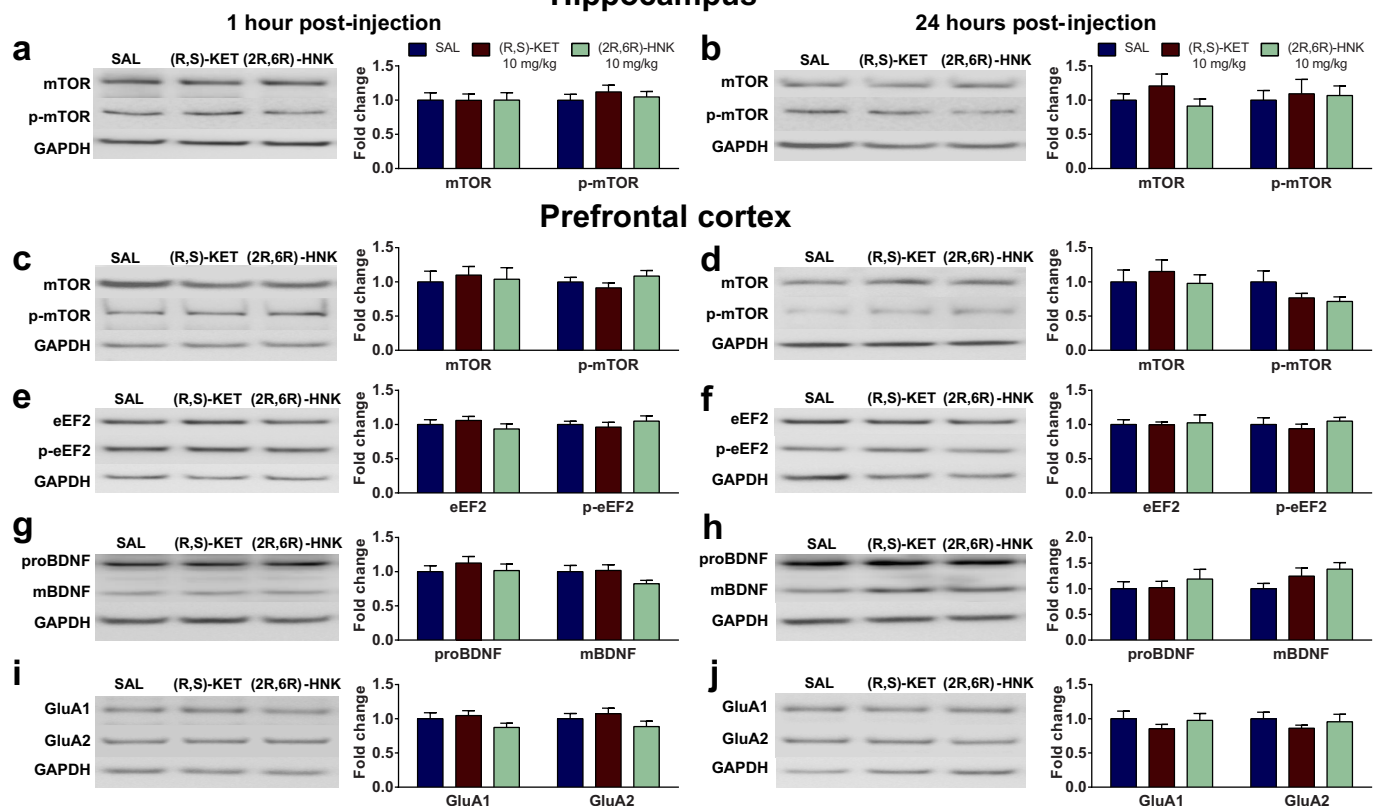
EEG - (R,S)-KET vs (2R,6R)-HNK



Extended Data Figure 7 | Administration of the AMPAR antagonist, NBQX, prevents (2R,6R)-HNK-induced increases in gamma oscillations *in vivo*. a, Administration of (R,S)-KET, but not (2R,6R)-HNK, increased locomotor home-cage activity of mice. b–e, Neither (R,S)-KET nor (2R,6R)-HNK altered cortical alpha (b), beta (c), delta (d) or theta (e) oscillations *in vivo*. f–k, Pre-treatment with the AMPAR

antagonist, NBQX, did not change the locomotor activity (f), alpha (g), beta (h), delta (j) or theta (k) oscillations, but it prevented (2R,6R)-HNK-induced increases of gamma oscillations *in vivo* (i). Data are mean \pm s.e.m. * $P < 0.05$, ** $P < 0.01$ (see Supplementary Table 1 for statistical analyses and n numbers).

Hippocampus

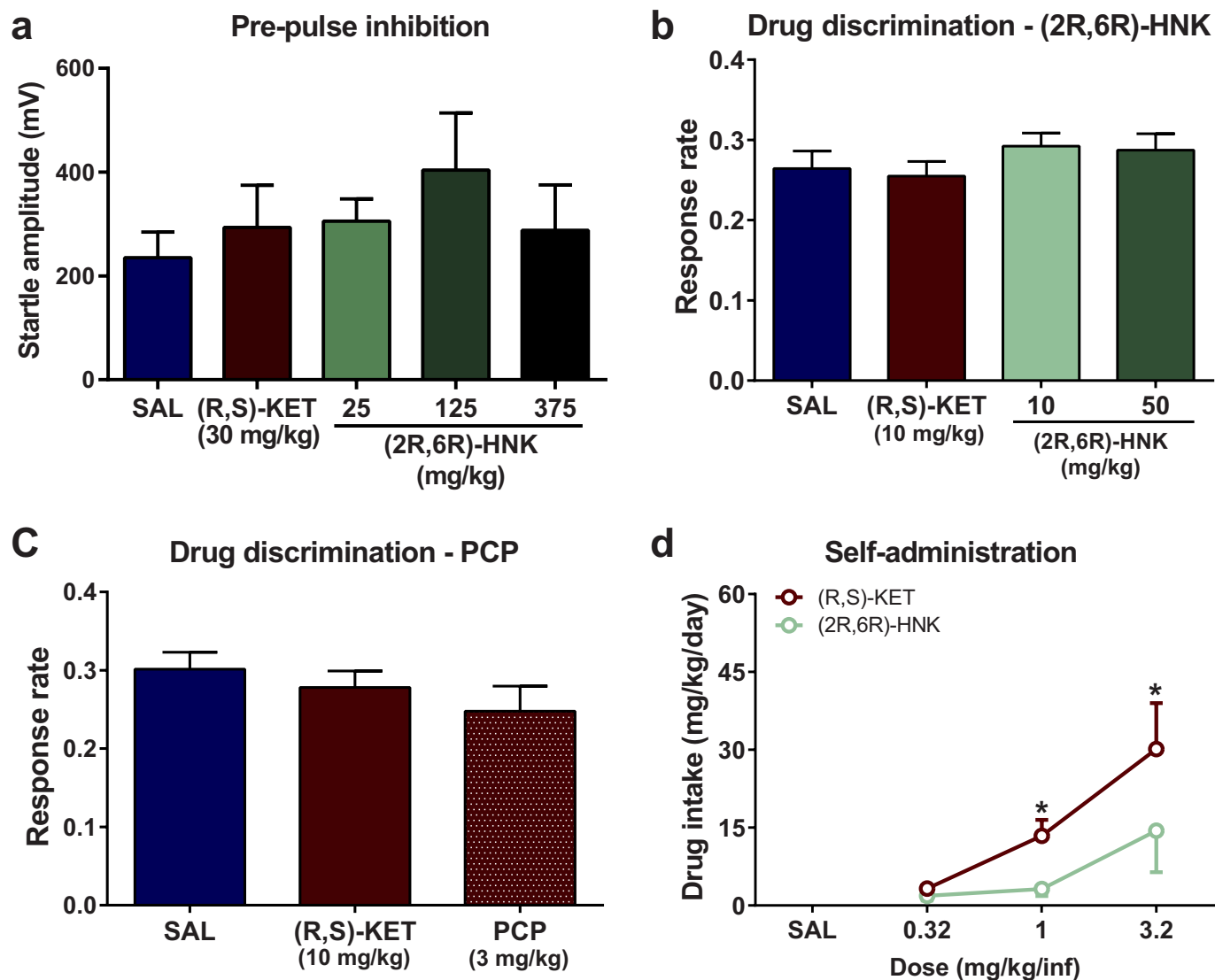


Extended Data Figure 8 | Effects of (2R,6R)-hydroxynorketamine on protein and protein phosphorylation levels in synaptoneurosomes.

a, b, A single administration of (R,S)-KET (10 mg kg⁻¹) or (2R,6R)-HNK (10 mg kg⁻¹), did not alter levels of mTOR or phosphorylated mTOR in the hippocampus 1 h (**a**) or 24 h (**b**) after injection.

c–j, Administration of (R,S)-KET or (2R,6R)-HNK did not alter levels of mTOR/phosphorylated mTOR (**c, d**), eEF2/phosphorylated eEF2 (**e, f**), proBDNF/mBDNF (**g, h**), or GluA1/GluA2 (**i, j**), in the prefrontal

cortex of mice. The values for the phosphorylated forms of proteins were normalized to phosphorylation-independent levels of the same protein. Phosphorylation-independent levels of proteins were normalized to GAPDH. Data are mean ± s.e.m, and were normalized to the saline-treated control group for each protein. Images are cropped; see Supplementary Fig. 1 for complete blot images. **P* < 0.05 (see Supplementary Table 1 for statistical analyses and *n* numbers).



Extended Data Figure 9 | (2R,6R)-HNK administration does not alter startle amplitude, drug discrimination rate or self-administration drug intake. **a**, Startle amplitude as measured in the pre-pulse inhibition task was not affected by administration of (R,S)-KET or (2R,6R)-HNK. **b**, **c**, Response rate of overall lever pressing per sec in the drug

discrimination model was not changed by administration of (R,S)-KET, (2R,6R)-HNK (**b**) or PCP (**c**). **d**, Unlike ketamine, (2R,6R)-HNK did not alter drug intake in the self-administration task in mice. Data are mean \pm s.e.m. * $P < 0.05$ (see Supplementary Table 1 for statistical analyses and n numbers).

Tracing haematopoietic stem cell formation at single-cell resolution

Fan Zhou^{1*}, Xianlong Li^{2*}, Weili Wang^{3*}, Ping Zhu^{2,4*}, Jie Zhou^{1*}, Wenyan He^{1*}, Meng Ding¹, Fuyin Xiong¹, Xiaona Zheng¹, Zhuan Li¹, Yanli Ni¹, Xiaohuan Mu³, Lu Wen^{2,5}, Tao Cheng^{3,6}, Yu Lan⁷, Weiping Yuan³, Fuchou Tang^{2,4,5,8} & Bing Liu^{1,3,9}

Haematopoietic stem cells (HSCs) are derived early from embryonic precursors, such as haemogenic endothelial cells and pre-haematopoietic stem cells (pre-HSCs), the molecular identity of which still remains elusive. Here we use potent surface markers to capture the nascent pre-HSCs at high purity, as rigorously validated by single-cell-initiated serial transplantation. Then we apply single-cell RNA sequencing to analyse endothelial cells, CD45[−] and CD45⁺ pre-HSCs in the aorta-gonad-mesonephros region, and HSCs in fetal liver. Pre-HSCs show unique features in transcriptional machinery, arterial signature, metabolism state, signalling pathway, and transcription factor network. Functionally, activation of mechanistic targets of rapamycin (mTOR) is shown to be indispensable for the emergence of HSCs but not haematopoietic progenitors. Transcriptome data-based functional analysis reveals remarkable heterogeneity in cell-cycle status of pre-HSCs. Finally, the core molecular signature of pre-HSCs is identified. Collectively, our work paves the way for dissection of complex molecular mechanisms regulating stepwise generation of HSCs *in vivo*, informing future efforts to engineer HSCs for clinical applications.

Haematopoietic stem cells (HSCs) are capable of both self-renewing extensively and differentiating progressively into entire mature blood lineages. The first HSC appears in mid-gestational mouse embryos at embryonic day 10.5 (E10.5)^{1–5}. From E12, HSCs progressively colonize fetal liver and dramatically expand there. Whereas embryonic stem cells injected into blastocysts can easily develop into normal animals with homeostatic HSCs, *in vitro* induction of such pluripotent stem cells into functional HSCs remains largely unsuccessful, reflecting much limited understanding of HSC ontogeny *in vivo*^{6,7}. Besides haemogenic endothelial cells (ECs), at least two HSC-competent intermediates around E11 have been revealed as precursors of HSCs, namely the pre-haematopoietic stem cells (pre-HSCs)^{8,9}. The emergence of CD45[−] pre-HSCs (type 1 pre-HSCs, referred to as T1 pre-HSCs) indicates the segregation from endothelial lineage by priming CD41 expression and losing endothelial potential concurrently^{9,10}. Subsequently, they proceed further into CD45⁺ pre-HSCs (type 2 pre-HSCs, T2 pre-HSCs), which ultimately form mature HSCs^{9,11}. Unlike mature HSCs that can directly repopulate irradiated recipients, identification of pre-HSCs must combine the steps of initial culture induction and the following transplantation^{2,9,12}. Sitting at the pinnacle of haematopoietic hierarchy, HSCs are rare, constituting fewer than 1 in 10,000 fetal liver and adult bone marrow cells. However, HSCs can be prospectively isolated to high purities by using antibody cocktails, thereby ensuring subsequent manipulations at the single-cell level^{13–18}. Unfortunately, this is not the case for HSC-competent cells in mouse mid-gestation embryos.

Recently, we and other groups have developed single-cell RNA-sequencing (RNA-seq) techniques that can analyse the transcriptome at single-cell and single-base resolutions¹⁹. These techniques have greatly facilitated the dissection of gene expression networks in rare cell types, such as the blastomeres of human early embryos²⁰,

and promoted the dissection of gene expression heterogeneity within temporally and spatially complex cell populations^{21–23}.

Here we use a novel combination of surface markers to isolate individual functional CD45[−] and CD45⁺ pre-HSCs in the E11 mouse aorta-gonad-mesonephros (AGM) region efficiently. By single-cell RNA-seq analyses, we reveal remarkable heterogeneity in relevant populations along with HSC generation. We further demonstrate a unique role of endothelial mTOR complex 2 (mTORC2) signalling in HSC formation during mouse embryogenesis. Our RNA-seq data and analysis results are accessible and can be explored for haematopoiesis research at <http://www.singlecell.pku.edu.cn/HSC>.

Isolation of single T1 pre-HSC by CD201

T1 pre-HSCs are known as CD31⁺VE-cadherin⁺CD45[−]CD41^{low}c-Kit⁺ in E11 AGM^{9,12,24}. To further enrich the population, we evaluated the efficacy of several candidate surface markers including VE-cadherin, AA4.1, and CD201 (also named endothelial protein C receptor, EPCR)^{16,25,26} (Fig. 1a and Extended Data Fig. 1a–c). Co-culture plus transplantation revealed that robust short-term reconstitution occurred only in CD201^{high} but not in CD201^{low/−} groups within the CD31⁺CD45[−]CD41^{low}c-Kit⁺ population in the E11 AGM region (Fig. 1b, c). To further functionally quantify the CD31⁺CD45[−]CD41^{low}c-Kit⁺CD201^{high} population, we performed co-cultures seeded with three cells and single cells, respectively (Fig. 1d), followed by transplantation. Of the recipients, 81.0% and 31.3% showed apparent long-term multi-lineage reconstitution, respectively (Fig. 1e–h and Supplementary Table 1). Secondary transplantation further confirmed the self-renewal ability of donor HSCs from three primary repopulated recipients receiving single-cell-initiated cultures (Fig. 1i and Extended Data Fig. 1d). These *in vivo* functional data validated that the CD31⁺CD45[−]CD41^{low}c-Kit⁺CD201^{high} subset in the E11 AGM

¹State Key Laboratory of Proteomics, Translational Medicine Center of Stem Cells, 307-Ivy Translational Medicine Center, Laboratory of Oncology, Affiliated Hospital, Academy of Military Medical Sciences, Beijing 100071, China. ²Biodynamic Optical Imaging Center, College of Life Sciences, Peking University, Beijing 100871, China. ³State Key Laboratory of Experimental Hematology, Institute of Hematology and Blood Diseases Hospital, Chinese Academy of Medical Sciences, Tianjin 300020, China. ⁴Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China. ⁵Ministry of Education Key Laboratory of Cell Proliferation and Differentiation, Beijing 100871, China. ⁶Collaborative Innovation Center for Cancer Medicine, National Institute of Biological Sciences, Tianjin 300020, China. ⁷State Key Laboratory of Proteomics, Genetic Laboratory of Development and Diseases, Institute of Biotechnology, Beijing 100071, China. ⁸Center for Molecular and Translational Medicine (CMTM), Beijing 100101, China. ⁹Institute of Hematology, Medical College of Jinan University, Guangzhou 510632, China.

*These authors contributed equally to this work.

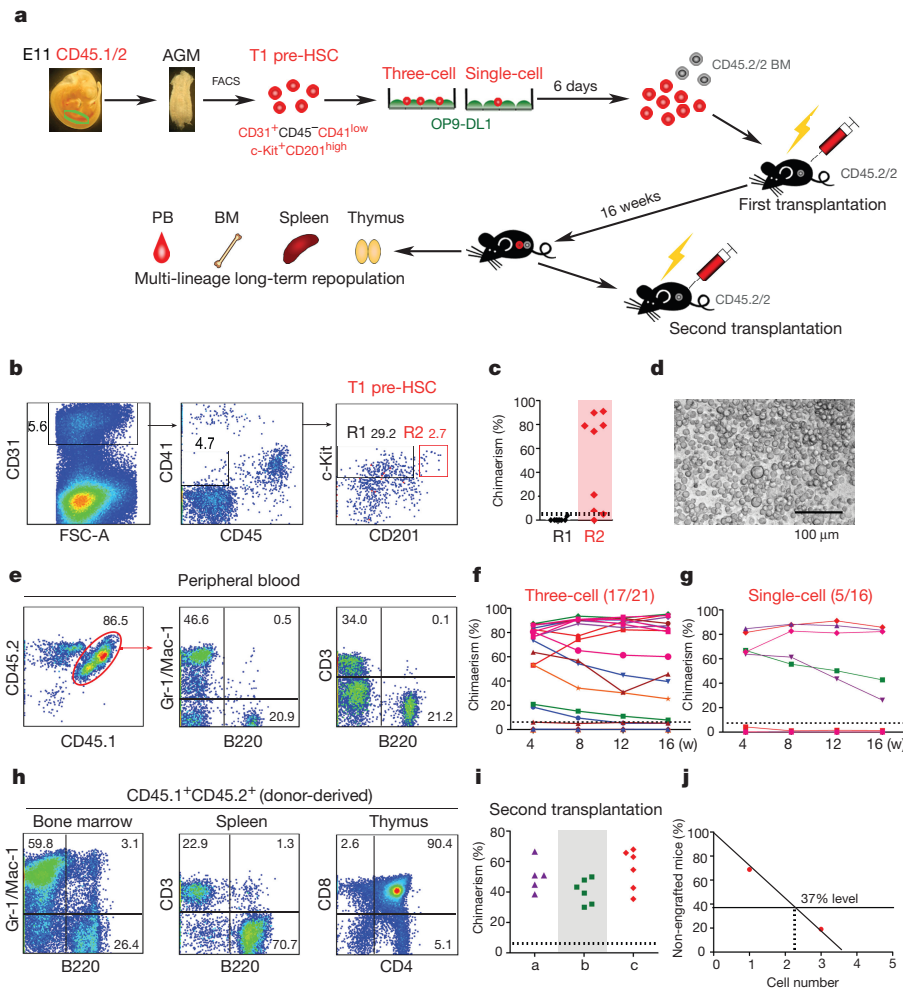


Figure 1 | Identification of T1 pre-HSCs at single-cell resolution.

a, Schematic illustration of the strategy. PB, peripheral blood; BM, bone marrow. **b**, Enrichment of T1 pre-HSCs according to CD201 expression. **c**, Donor chimaerism in peripheral blood of recipients 4 weeks after transplantation. **d**, Morphology of haematopoietic progenies generated by a single T1 pre-HSC after OP9-DL1 co-culture. **e**, Multi-lineage long-term (>16 weeks) repopulation in peripheral blood of a representative primary recipient transplanted with the culture from a single T1 pre-HSC. Donor-derived (CD45.1⁺CD45.2⁺) myeloid (Gr-1⁺/Mac-1⁺), B-lymphoid (B220⁺), and T-lymphoid (CD3⁺) cells are shown. **f, g**, Donor chimaerism in peripheral blood of primary recipients in three-cell and single-cell

groups monitored at 4, 8, 12, and 16 weeks after transplantation. Numbers within brackets indicate successfully reconstituted recipients/number transplanted. **h**, Representative multi-lineage long-term (>16 weeks) repopulation in haematopoietic organs of a primary recipient of a single T1 pre-HSC. **i**, Repopulating potential of bone marrow cells from three repopulated primary recipients (a, b and c) of single T1 pre-HSC group. Symbols represent donor chimaerism in peripheral blood of 17 secondary recipients 12 weeks after transplantation. **j**, Quantification of T1 pre-HSC frequency by limiting dilution analysis. Detailed information related to all the transplantation assays in this study is available in Supplementary Information.

region (0.007%, approximately 11 per tissue) highly enriched T1 pre-HSCs at a frequency of 1:2.3, as calculated by limiting dilution analyses (Fig. 1j). Therefore, the highly purified CD45⁺ T1 pre-HSC population enabled, for the first time, further analyses of the elusive nascent HSC precursor at single-cell resolution.

T2 identification by single-cell RNA-seq

We also determined the inducible HSC potential from the presumed CD45⁺ T2 pre-HSCs in the E11 AGM region. The CD31⁺CD45⁺CD41^{low} (T2 CD41^{low}) population demonstrated long-term multi-lineage repopulation in 37.5% of recipients after three-cell-initiated co-culture (Extended Data Fig. 1e–h and Supplementary Table 1), but failed to reconstitute the recipients with direct transplantation (0 of 9, 28–43 cells from 1.5 to 2.0 embryo equivalents per recipient).

Five populations related to HSC ontogeny were prepared: in the E11 AGM region, (1) ECs, CD31⁺VE-cadherin⁺CD41⁺CD43⁺CD45⁺Ter119⁺; (2) T1 pre-HSCs, CD31⁺CD45⁺CD41^{low}c-Kit⁺CD201^{high}; (3) T2 CD41^{low}, CD31⁺CD45⁺CD41^{low}; and in fetal liver, (4) E12 HSCs, Lin[−]Sca-1⁺Mac-1^{low}CD201⁺; and (5) E14 HSCs,

CD45⁺CD150⁺CD48[−]CD201⁺ (ref. 16; Fig. 2a). Both ten-cell and single-cell RNA-seq analyses were performed for each of the five cell types (Extended Data Figs 2 and 3a–d and Supplementary Tables 2–4). Of note, principal component analysis (PCA) and unsupervised hierarchical clustering analyses of single-cell RNA-seq data revealed that the T2 CD41^{low} population was clearly separated into two distinct subpopulations, with subpopulation A showing similar gene expression profiles to T1 pre-HSCs (Fig. 2b, c and Extended Data Fig. 3d). Although the two subpopulations similarly expressed the genes used for sorting (Extended Data Fig. 3e), they showed mutually exclusive expression of several molecules (Fig. 2c and Extended Data Fig. 3f). Subpopulation A consistently expressed several EC markers (including *Procr*, the coding gene for CD201) and HSC-related transcription factors. However, subpopulation B possessed myeloid lineage signatures. The heterogeneity within the T2 CD41^{low} population suggested myeloid cell contamination in the presumed T2 pre-HSCs. Subsequently, we sorted the CD31⁺CD45⁺c-Kit⁺CD201^{high} population and performed co-culture/transplantation assays (Fig. 2a and Extended Data Fig. 3g). The functional data showed that

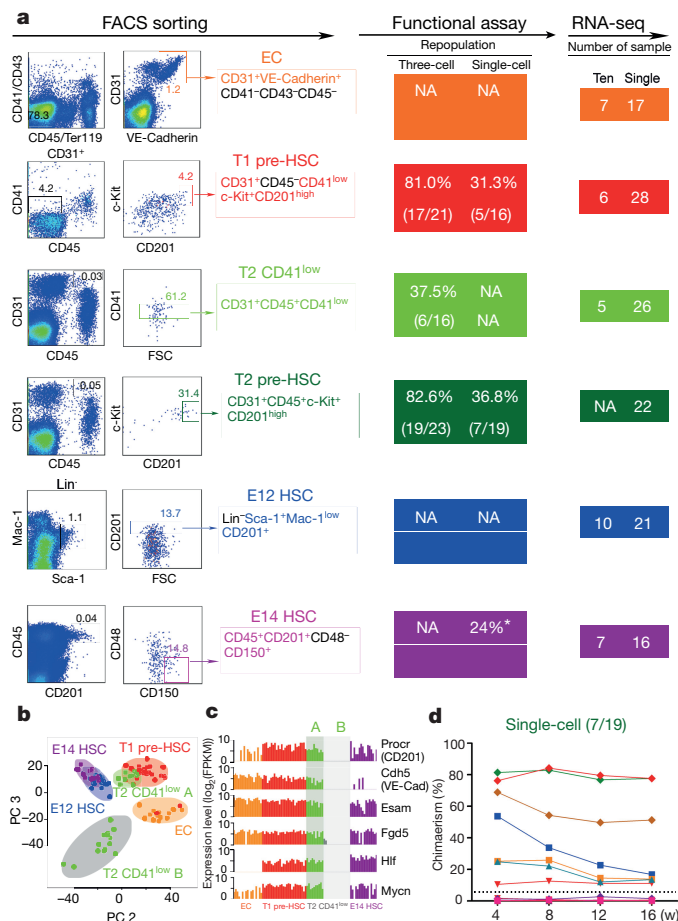


Figure 2 | Enhanced enrichment of T2 pre-HSCs by single-cell RNA-seq. **a**, Schematic illustration of fluorescence-activated cell sorting (FACS) preparation of six types of cells (left), donor chimaerism of three-cell- or single-cell-initiated culture/transplantation (middle), and number of RNA-seq sample (right). *Repopulation efficiency is taken from ref. 16. **b**, PCA of 108 single cells. **c**, Selected markers with different expression between A and B subpopulations of T2 CD41^{low}. **d**, Donor chimaerism in peripheral blood of recipients of single T2 pre-HSCs monitored at the indicated weeks after transplantation. Numbers within brackets indicate successfully reconstituted recipients/number transplanted (a and d). NA, not available.

the CD31⁺CD45⁺c-Kit⁺CD201^{high} subset in the E11 AGM region (0.012%, approximately 18 per tissue) highly enriched T2 pre-HSCs at a frequency of 1:2.1 as calculated by limiting dilution analyses (Fig. 2d, Extended Data Fig. 3h and Supplementary Table 5). Finally, we performed single-cell RNA-seq of 22 CD31⁺CD45⁺c-Kit⁺CD201^{high} cells, which were used as T2 pre-HSCs for subsequent transcriptome analyses.

Global gene expression dynamics

We excluded five cells that mixed with other populations and kept 99 cells in the single-cell RNA-seq data for further analyses (Extended Data Fig. 4a). Unsupervised hierarchical clustering, PCA, and *t*-distributed stochastic neighbour embedding analyses revealed three clusters: ECs, pre-HSCs, and mature HSCs. The T1 and T2 pre-HSCs were relatively similar to each other whereas E12 and E14 HSCs were partly mixed with each other (Fig. 3a and Extended Data Fig. 4a, b). We also performed pseudo-time analysis by the Monocle method²⁷, and the pseudo-time developmental path gave the same order as that by PCA, showing continuous development from ECs to HSCs through T1 and T2 pre-HSCs (Extended Data Fig. 4c).

On average, each EC expressed 0.32 million copies of mRNAs, much lower than 1.41 million in T1 pre-HSCs (Fig. 3b). This was represented

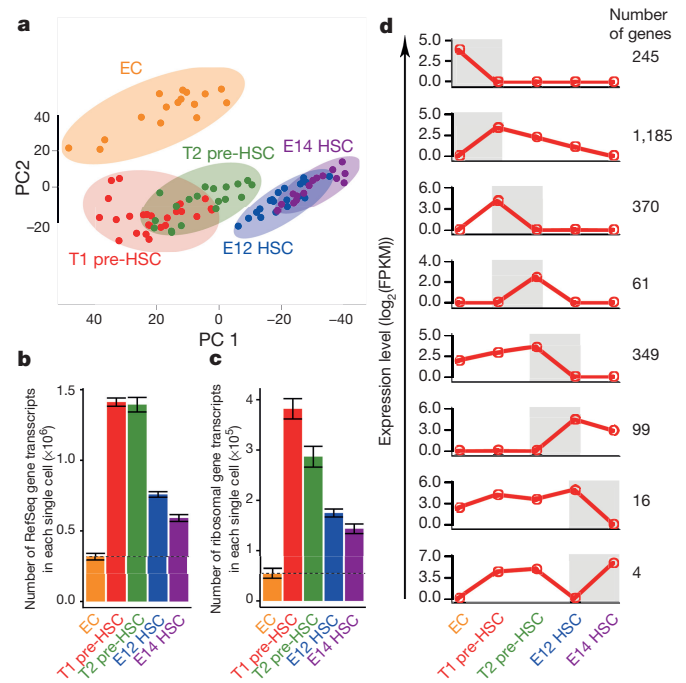


Figure 3 | Global gene expression dynamics during HSC formation. **a**, PCA of 99 single cells. **b**, **c**, Average total copy numbers of the mRNAs (b) and ribosome-associated mRNAs (c) in an individual cell of each cell type (error bar, mean $\pm 2 \times$ s.e.m.). The dashed line shows the equivalence of average value in ECs. **d**, Dynamic changes of differentially expressed genes between each of the two consecutive stages, with the number of differentially expressed genes indicated to the right.

by a sharp increase in ribosome-based translational machinery, as the absolute expression levels of 290 ribosome-associated genes in T1 pre-HSCs increased by 6.9-fold (Fig. 3c). In contrast, during HSC maturation from pre-HSCs, the transcriptional activity in each individual HSC was gradually reduced (Fig. 3b).

There were 2,060 genes showing significant changes between neighbouring cell types (on the basis of multiple *t*-tests with $P < 0.05$ and false discovery rate (FDR) < 0.05 and fold-changes of log₂-converted fragment per kilobase of transcript per million fragments mapped (FPKM) > 2 or < 0.5); these genes formed eight clusters (Fig. 3d, Extended Data Fig. 4d and Supplementary Table 6). Gene ontology (GO) analysis was used to determine enriched terms (Supplementary Table 7). The most dramatic change occurred between ECs and T1 pre-HSCs. The genes downregulated clearly enriched for terms related to cell migration, vasculature development, and blood vessel morphogenesis, while the genes upregulated strongly enriched for haematopoietic or lymphoid organ development and intracellular signalling cascade (Extended Data Fig. 4d).

HSCs are derived from embryonic ECs, but the endothelial signature of pre-HSCs remains undefined. Here we analysed 324 genes from the blood-vessel-related GO terms. Most of the differentially expressed genes were those that were gradually downregulated from ECs to HSCs (Extended Data Fig. 4e and Supplementary Table 6). From ECs to pre-HSCs, the numbers of expressed angiogenesis genes maintained but their total expression levels in each individual cell gradually decreased (Extended Data Fig. 4f, g). Supportively, pre-HSCs have been reported to lose endothelial potential *in vitro*⁹. It is believed that only ECs located in major arteries but not in veins give rise to HSCs. Recent studies suggest that arterial ECs and HSCs originate from distinct precursors^{28,29}. We examined the expression patterns of artery- and vein-specific genes during HSC specification. At the individual cell level, the arterial and venous features were segregated in E11 AGM-derived ECs. T1 pre-HSCs expressed evident arterial markers but much lower levels of venous markers compared with ECs

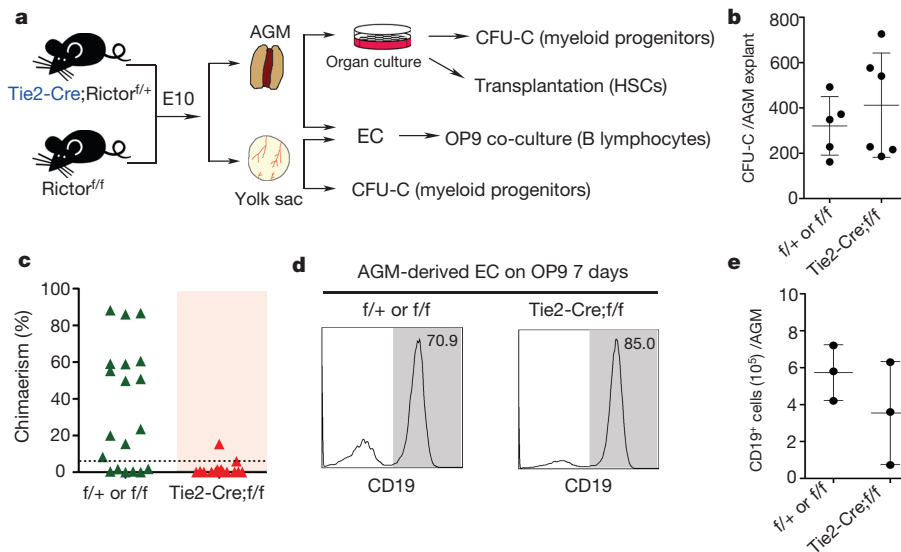


Figure 4 | Specific role of mTORC2 signalling during HSC formation. **a**, Schematic experimental design. **b**, Quantification of CFU-Cs in E10 AGM (31–35 somite pairs) of *Tie2-Cre;Rictor^{f/f}* and littermate control embryos after 72-h organ culture. Data are collected from five (*f/+* or *f/f*) and six (*Tie2-Cre;f/f*) CFU-C cultures. **c**, Donor chimaerism in peripheral blood of recipients 16 weeks after transplantation. Data are collected from

19 and 15 recipients transplanted with a total of 43 (*f/+* or *f/f*) and 33 (*Tie2-Cre;f/f*) embryos in 6 independent experiments. **d**, **e**, B lymphoid potential of ECs (CD31⁺CD41[−]CD45[−]Ter119[−]) from E10 AGM after OP9 co-culture. Data are collected from *n* = 6 embryos per genotype over three independent experiments. Error bar, mean \pm s.d.

(Extended Data Fig. 4h and Supplementary Table 8), suggesting that pre-HSCs should have a more intimate lineage relationship with arterial ECs than with venous ECs. Supportively, there might be a common endothelial precursor that is specified to the arterial or haemogenic lineage during embryogenesis²⁹.

Specific role of mTORC2 signalling

Next, the gene set enrichment analysis (GSEA) coupled with the pathway gene set data of Kyoto Encyclopedia of Genes and Genomes (KEGG) was applied by pairwise comparison to explore signalling pathways potentially involved in HSC formation (Extended Data Fig. 5a, b). The result by GSEA was verified by hypergeometric testing based on differentially expressed genes. Compared with ECs, there were 57 pathways overrepresented in T1 pre-HSCs ($P < 0.05$, FDR < 0.25 ; Supplementary Table 9). Notably, oxidative phosphorylation and glycolysis pathways were both overrepresented in the T1 population, implying that mitochondrial aerobic respiration was specifically activated at this stage. In contrast, quiescent HSCs reside in hypoxic bone marrow niches and rely heavily upon glycolysis for energy production³⁰.

Additionally, we found that the mTOR signalling pathway was highly enriched in T1 pre-HSCs compared with ECs. mTOR consists of two multi-protein complexes: mTOR complex 1 (mTORC1) and mTOR complex 2 (mTORC2)³¹. Higher expression of growth factor receptors (Fgfr3, S1pr4, and Tfr2), robust ribosomal subunits, and enhanced membrane-cytoskeleton-coupled cellular processes were observed in T1 pre-HSCs, indicative of mTORC2 activation (Extended Data Fig. 5b and Supplementary Table 6).

As a core mTORC2 component, Rictor is unnecessary for maintenance and function of HSCs but pivotal for B-cell lineage differentiation^{32,33}. Here we disrupted Rictor from the embryonic endothelial stage to determine whether mTORC2 is required for HSC emergence from endothelium. The *Tie2-Cre;Rictor^{f/f}* embryos displayed gross phenotypes including scattered haemorrhage and growth delay from E10.5 (35–40 somite pairs); thus we did not choose this stage onwards for haematopoietic assay (Extended Data Fig. 6a). We used an organ culture strategy that could facilitate robust HSC formation within AGM tissues *in vitro* as early as 30 somite pairs³⁴ (Fig. 4a). After organ culture, the *Tie2-Cre;Rictor^{f/f}* AGM explants showed normal cell number, cell viability, CD45⁺c-Kit⁺ percentage, and CFU-C number, but

strikingly lower repopulating capacity than controls (Fig. 4b, c and Extended Data Fig. 6b–f). We further examined the haematopoietic progenitor potential including B-lymphocyte and myeloid lineages, and found both were not significantly lower in *Tie2-Cre;Rictor^{f/f}* group compared with the controls (Fig. 4d, e and Extended Data Fig. 6g–i). The results indicated that the endothelial perturbation of mTORC2 signalling specifically dampened the generation of HSCs but not haematopoietic progenitor cells (HPCs). We further used the *Vav-Cre* transgenic mice to delete the Rictor gene from embryonic haematopoietic cells. In the E12.5 *Vav-Cre;Rictor^{f/f}* AGM, we detected functional HSCs showing multi-lineage reconstitution despite reduced chimaerism of donor cells by direct transplantation (Extended Data Fig. 6j–l). Therefore, Rictor was indispensable for HSC emergence from endothelial cells, but was required less in later haematopoietic progression during development.

Dynamic transcription factor and surface marker expression

We explored the pattern genes, namely the developmental-stage as well as the cell-type specifically expressed genes, during HSC formation using PCA. We found four dominant expression patterns by comparing ECs and pre-HSCs and five patterns by comparing pre-HSCs and HSCs (Extended Data Fig. 7a, b). Transcription factors are crucial to orchestrate cell fate transition. We analysed the dynamic transcription factor network within the pattern genes between ECs and pre-HSCs (Extended Data Fig. 7c, d). Some transcription factors essential for haematopoietic specification from mesodermal precursors were detected in all five populations (Extended Data Fig. 7e). In contrast, several HSC-related transcription factors, including the master regulator Runx1 and its targets, were detected in none of the ECs but in all T1 pre-HSCs (Extended Data Fig. 7f). Most heptad transcription factors³⁵ were abundantly expressed in pre-HSCs and HSCs (Extended Data Fig. 8a). Albeit at varying frequencies and levels, the T1 pre-HSCs and HSCs expressed all eight transcription factors, the combination of which could confer committed progenitors with HSC potential³⁶ (Extended Data Fig. 8b). Several inflammation-related genes showed higher expression in T1 pre-HSCs compared with ECs (Extended Data Fig. 8c), supportive of a recently established correlation between inflammatory signals and HSC emergence in vertebrates^{37,38}.

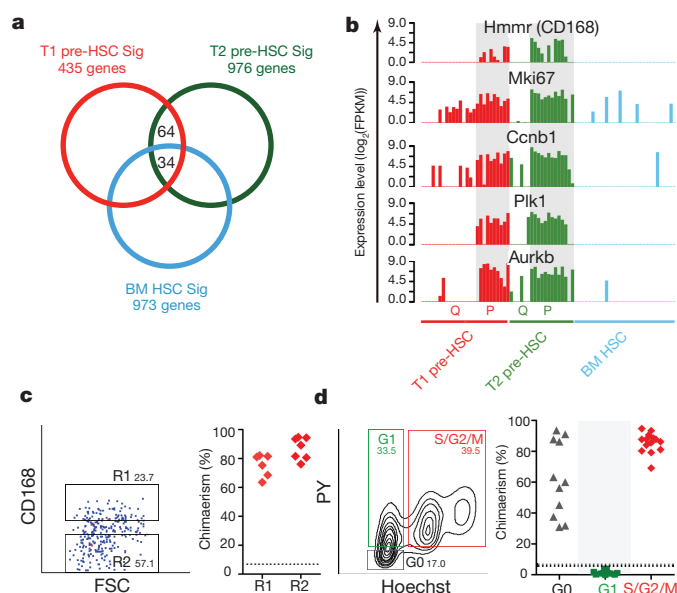


Figure 5 | Signature genes and cell-cycle heterogeneity of pre-HSCs.

a, Venn diagram shows shared and distinct signature genes among T1 pre-HSCs, T2 pre-HSCs, and bone marrow HSCs. Sig, signature. **b**, Expression of selected cell-division-related genes in pre-HSCs and adult HSCs. **c**, Repopulating potential of CD168⁺ (R1) and CD168[−] (R2) T1 pre-HSCs (CD31⁺CD45[−]CD41^{low}). Donor chimaerism in peripheral blood of recipients 8 weeks after transplantation is shown to the right. **d**, Repopulating potential of T1 pre-HSCs (CD31⁺CD45[−]) at G0 (black), G1 (green), and S/G2/M phases (red), respectively. Donor chimaerism in peripheral blood of recipients 3 weeks after transplantation is shown to the right. The numbers of recipients are $n = 6$ (R1) and $n = 7$ (R2) for **c**; $n = 11$ (G0), $n = 11$ (G1), and $n = 14$ (S/G2/M) for **d**.

Given the five populations defined according to surface markers, we analysed their epitope features using 264 recently reported candidate molecules¹⁸. A few molecules, such as Procr (CD201), Kit, Cd47, Cd55, and Cd63, were expressed in a fraction of ECs and were homogeneously expressed in pre-HSCs and HSCs, implying their presumably continuous marking of the HSC formation process (Fig. 2c and Extended Data Fig. 8d). Then Cd47 was selected to address this issue, which was expressed in approximately half of putative ECs at E10, the earliest time point when HSC-competent ECs and CD45[−] pre-HSCs could be detected in AGM region^{9,12}. Upon induction by OP9-DL1 co-culture, only cells derived from the CD47⁺ subset both of ECs and of CD45[−] pre-HSCs could significantly long-term multi-lineage engraft irradiated recipients (Extended Data Fig. 8e). Using direct transplantation, long-term chimaerism was detected in the c-Kit⁺CD47⁺ but not the c-Kit⁺CD47[−] subset of the E11.5 AGM region (Extended Data Fig. 8f, g). This suggested that the HSC-competent ECs, CD45[−] pre-HSCs, and mature HSCs were exclusively positive for CD47 in the E10–E11 AGM region (Supplementary Table 10).

Long non-coding RNAs (lncRNAs) have recently been shown as important for numerous biological processes³⁹. We also analysed the expression of known lncRNAs during HSC specification based on the GENCODE lncRNA database (Extended Data Fig. 9a–e and Supplementary Table 11). On average, T2 pre-HSCs expressed most abundant amount of lncRNAs among five populations (Extended Data Fig. 9a). Among the 26 lncRNAs specifically expressed in adult bone-marrow-derived HSCs⁴⁰, only H19 and Malat1 were expressed significantly in most HSC-related cells (Extended Data Fig. 9f).

Signature genes and cell-cycle feature

The molecular signature of pre-HSCs was obtained by comparing them with closely related populations without pre-HSC potential. The T1 pre-HSC signature genes indicated the 435 genes specifically

expressed in T1 pre-HSCs (R2 population in Fig. 1b) but not in the CD31⁺CD45[−]CD41^{low}c-Kit⁺CD201^{low/−} population (R1 population in Fig. 1b) (Fig. 5a and Supplementary Table 12). Similarly, the T2 pre-HSC signature genes indicated the 976 genes highly enriched in the functional T2 pre-HSCs (CD31⁺CD45⁺CD41^{low}CD201⁺ population A in Fig. 2b; and CD31⁺CD45⁺c-Kit⁺CD201^{high}) but not the CD31⁺CD45⁺CD41^{low}CD201[−] population with myeloid features (population B in Fig. 2b) (Fig. 5a and Supplementary Table 12). The 98 genes that were shared between T1 and T2 pre-HSCs were designated as pre-HSC signature genes (Fig. 5a, Extended Data Fig. 10a and Supplementary Table 12). As a component preserving mTOR signalling pathway⁴¹, Eya2 was also enriched, confirming the role of the mTORC2 pathway in HSC emergence. Moreover, 30% of the pre-HSC signature genes were shared by adult bone marrow HSCs reported in the literature^{40,42–44}. The 98 pre-HSC signature genes were mainly enriched in ‘positive regulation of developmental process’, ‘regulation of transcription’, and ‘intracellular signalling cascade’ (Extended Data Fig. 10b). Network analysis of transcription factors within the pre-HSC signature genes was also performed (Extended Data Fig. 10c). Interestingly, one of the pre-HSC signature genes, *Ctnnal1* (or *a-catulin*), was recently reported to efficiently mark adult bone marrow HSCs⁴⁵. Therefore, most of the pre-HSC signature genes deserve comprehensive analyses and evaluation in haematopoietic systems both of embryos and of adults.

In contrast to adult HSCs in a consistently static state as expected, the pre-HSC populations showed pronounced divergence in the expression pattern of 304 cell-division-related genes, suggesting the existence of both actively proliferative and relatively quiescent states (Extended Data Fig. 10d). Among the 304 genes, *Hmhr* (coding gene for CD168) was the only surface marker and was expressed in most of the actively cycling pre-HSCs but none of the rest quiescent pre-HSCs or adult bone marrow HSCs (Fig. 5b), in line with previous reports of higher CD168 expression in the cells of G2/M phase⁴⁶. Flow cytometry analyses further confirmed the actively proliferating state of CD168⁺ cells and a limited proliferation state of CD168[−] cells within the CD31⁺CD45[−]CD41^{low} population (Extended Data Fig. 10e). Functionally, T1 pre-HSCs resided both in CD168⁺ and in CD168[−] subpopulations (Fig. 5c).

Furthermore, we analysed the cell-cycle status of functional T1 pre-HSCs. The cells in S/G2/M phases exhibited the most remarkable proliferation in co-cultures and the most notable reconstitution potential. The T1 cells in G0 phase showed less robust proliferation *in vitro* and successful repopulation with lower chimaerism. In striking contrast, none of the recipients were repopulated by T1 cells in G1 phase, which hardly proliferated in the co-cultures (Fig. 5d). Consistently, the single-cell co-cultures of immunophenotypically defined pre-HSCs with high proliferation capacity showed the most remarkable reconstitution, in line with the previous finding that the cells without proliferation capacity *in vitro* cannot be functional pre-HSCs⁸ (Extended Data Fig. 10f). Of note, the significant S/G2/M phase population in functional pre-HSCs differs greatly from adult bone marrow HSCs, most of which belong to G0 phase, and from E14 fetal liver HSCs, all of which are in G1 phase⁴⁷, suggesting the stage-specific cell-cycle features during HSC development. Collectively, the results imply a previously unrecognized expansion of a proportion of pre-HSCs, before their maturation and migration to fetal liver.

Conclusion

To the best of our knowledge, this study is the first to successfully purify embryonic pre-HSCs at a single-cell resolution, 20 years later than adult HSCs¹⁴. The CD201^{high} feature was shown to be a potent marker recognizing both CD45[−] and CD45⁺ pre-HSCs. Recently, the putative precursor of pre-HSCs, termed pro-HSCs, was observed in E9.5 embryos, with a similar CD45[−]CD41^{low} immunophenotype but being much rarer than T1 pre-HSCs⁴⁸. The functional characterization of pro-HSCs at single-cell level deserves further efforts.

Importantly, this is the first time that the transcriptomes of pre-HSCs and HSCs during embryonic development have been comprehensively surveyed at single-cell and single-base resolution. The investigation of molecular profiles specific for HSC-competent populations is fundamental to uncovering novel mechanisms for HSC emergence. The HSCs and definitive HPCs are believed to arise from distinct subsets of haemogenic endothelium during development⁴⁹. Here we reveal the distinct regulatory strength of Rictor in the progression of vascular endothelial cells to HPCs and to HSCs, highlighting the complex and precise regulation in HSC emergence. In summary, our work paves the way for further dissection of the complex molecular regulation network concerning the ontogeny and maturation of HSCs, and for designing efficient strategies for HSC production in culture.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 September 2015; accepted 11 April 2016.

Published online 18 May 2016.

- Medvinsky, A. & Dzierzak, E. Definitive hematopoiesis is autonomously initiated by the AGM region. *Cell* **86**, 897–906 (1996).
- Zovein, A. C. *et al.* Fate tracing reveals the endothelial origin of hematopoietic stem cells. *Cell Stem Cell* **3**, 625–636 (2008).
- Dzierzak, E. & Speck, N. A. Of lineage and legacy: the development of mammalian hematopoietic stem cells. *Nature Immunol.* **9**, 129–136 (2008).
- Li, Z. *et al.* Mouse embryonic head as a site for hematopoietic stem cell development. *Cell Stem Cell* **11**, 663–675 (2012).
- Clements, W. K. & Traver, D. Signalling pathways that control vertebrate haematopoietic stem cell specification. *Nature Rev. Immunol.* **13**, 336–348 (2013).
- Murry, C. E. & Keller, G. Differentiation of embryonic stem cells to clinically relevant populations: lessons from embryonic development. *Cell* **132**, 661–680 (2008).
- Kyba, M., Perlingeiro, R. C. & Daley, G. Q. HoxB4 confers definitive lymphoid-myeloid engraftment potential on embryonic stem cell and yolk sac hematopoietic progenitors. *Cell* **109**, 29–37 (2002).
- Taoudi, S. *et al.* Extensive hematopoietic stem cell generation in the AGM region via maturation of VE-cadherin⁺CD45⁺ pre-definitive HSCs. *Cell Stem Cell* **3**, 99–108 (2008).
- Rybtsov, S. *et al.* Hierarchical organization and early hematopoietic specification of the developing HSC lineage in the AGM region. *J. Exp. Med.* **208**, 1305–1315 (2011).
- Ferkowicz, M. J. *et al.* CD41 expression defines the onset of primitive and definitive hematopoiesis in the murine embryo. *Development* **130**, 4393–4403 (2003).
- Yokomizo, T. & Dzierzak, E. Three-dimensional cartography of hematopoietic clusters in the vasculature of whole mouse embryos. *Development* **137**, 3651–3661 (2010).
- Li, Z. *et al.* Generation of hematopoietic stem cells from purified embryonic endothelial cells by a simple and efficient strategy. *J. Genet. Genomics* **40**, 557–563 (2013).
- Cheng, T. *et al.* Temporal mapping of gene expression levels during the differentiation of individual primary hematopoietic cells. *Proc. Natl Acad. Sci. USA* **93**, 13158–13163 (1996).
- Osawa, M., Hanada, K., Hamada, H. & Nakauchi, H. Long-term lymphohematopoietic reconstitution by a single CD34-low/negative hematopoietic stem cell. *Science* **273**, 242–245 (1996).
- Kiel, M. J. *et al.* SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell* **121**, 1109–1121 (2005).
- Benz, C. *et al.* Hematopoietic stem cell subtypes expand differentially during development and display distinct lymphopoietic programs. *Cell Stem Cell* **10**, 273–283 (2012).
- Etzrodt, M., Ende, M. & Schroeder, T. Quantitative single-cell approaches to stem cell research. *Cell Stem Cell* **15**, 546–558 (2014).
- Guo, G. *et al.* Mapping cellular hierarchy by single-cell analysis of the cell surface repertoire. *Cell Stem Cell* **13**, 492–505 (2013).
- Wang, Y. & Navin, N. E. Advances and applications of single-cell sequencing technologies. *Mol. Cell* **58**, 598–609 (2015).
- Yan, L. *et al.* Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature Struct. Mol. Biol.* **20**, 1131–1139 (2013).
- Durruthy-Durruthy, R. *et al.* Reconstruction of the mouse otocyst and early neuroblast lineage at single-cell resolution. *Cell* **157**, 964–978 (2014).
- Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).
- Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
- Kiessseian, A., Brunet de la Grange, P., Burlen-Defranoux, O., Godin, I. & Cumano, A. Immature hematopoietic stem cells undergo maturation in the fetal liver. *Development* **139**, 3521–3530 (2012).
- Iwasaki, H., Arai, F., Kubota, Y., Dahl, M. & Suda, T. Endothelial protein C receptor-expressing hematopoietic stem cells reside in the perisinusoidal niche in fetal liver. *Blood* **116**, 544–553 (2010).
- Godin, I., Garcia-Porrero, J. A., Dieterlen-Lièvre, F. & Cumano, A. Stem cell emergence and hemopoietic activity are incompatible in mouse intraembryonic sites. *J. Exp. Med.* **190**, 43–52 (1999).
- Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnol.* **32**, 381–386 (2014).
- Ditadi, A. *et al.* Human definitive haemogenic endothelium and arterial vascular endothelium represent distinct lineages. *Nature Cell Biol.* **17**, 580–591 (2015).
- Gama-Norton, L. *et al.* Notch signal strength controls cell fate in the haemogenic endothelium. *Nature Commun.* **6**, 8510 (2015).
- Wang, Y. H. *et al.* Cell-state-specific metabolic dependency in hematopoiesis and leukemogenesis. *Cell* **158**, 1309–1323 (2014).
- Laplanche, M. & Sabatini, D. M. mTOR signaling in growth control and disease. *Cell* **149**, 274–293 (2012).
- Magee, J. A. *et al.* Temporal changes in PTEN and mTORC2 regulation of hematopoietic stem cell self-renewal and leukemia suppression. *Cell Stem Cell* **11**, 415–428 (2012).
- Zhang, Y. *et al.* Rictor is required for early B cell development in bone marrow. *PLoS ONE* **9**, e103970 (2014).
- Robin, C. *et al.* An unexpected role for IL-3 in the embryonic development of hematopoietic stem cells. *Dev. Cell* **11**, 171–180 (2006).
- Wilson, N. K. *et al.* Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell* **7**, 532–544 (2010).
- Riddell, J. *et al.* Reprogramming committed murine blood cells to induced hematopoietic stem cells with defined factors. *Cell* **157**, 549–564 (2014).
- Espin-Palazón, R. *et al.* Proinflammatory signaling regulates hematopoietic stem cell emergence. *Cell* **159**, 1070–1085 (2014).
- Li, Y. *et al.* Inflammatory signaling regulates embryonic hematopoietic stem and progenitor cell production. *Genes Dev.* **28**, 2597–2612 (2014).
- Rinn, J. L. & Chang, H. Y. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81**, 145–166 (2012).
- Cabezas-Wallscheid, N. *et al.* Identification of regulatory networks in HSCs and their immediate progeny via integrated proteome, transcriptome, and DNA methylome analysis. *Cell Stem Cell* **15**, 507–522 (2014).
- Lee, S. H. *et al.* The transcription factor Eya2 prevents pressure overload-induced adverse cardiac remodeling. *J. Mol. Cell. Cardiol.* **46**, 596–605 (2009).
- Forsberg, E. C. *et al.* Differential expression of novel potential regulators in hematopoietic stem cells. *PLoS Genet.* **1**, e28 (2005).
- Chambers, S. M. *et al.* Hematopoietic fingerprints: an expression database of stem cells and their progeny. *Cell Stem Cell* **1**, 578–591 (2007).
- Gazit, R. *et al.* Transcriptome analysis identifies regulators of hematopoietic stem and progenitor cells. *Stem Cell Rep.* **1**, 266–280 (2013).
- Acar, M. *et al.* Deep imaging of bone marrow shows non-dividing stem cells are mainly perisinusoidal. *Nature* **526**, 126–130 (2015).
- Sohr, S. & Engeland, K. RHAMM is differentially expressed in the cell cycle and downregulated by the tumor suppressor p53. *Cell Cycle* **7**, 3448–3460 (2008).
- Bowie, M. B. *et al.* Hematopoietic stem cells proliferate until after birth and show a reversible phase-specific engraftment defect. *J. Clin. Invest.* **116**, 2808–2816 (2006).
- Rybtsov, S. *et al.* Tracing the origin of the HSC hierarchy reveals an SCF-dependent, IL-3-independent CD43(-) embryonic precursor. *Stem Cell Rep.* **3**, 489–501 (2014).
- Chen, M. J. *et al.* Erythroid/myeloid progenitors and hematopoietic stem cells originate from distinct populations of endothelial cells. *Cell Stem Cell* **9**, 541–552 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank H. Wu, J. Zheng, and H. Guo for discussion, and S. Hou and L. Zhang for technique support. This work was supported by the Chinese National Key Program on Basic Research (2011CB964800, 2012CB966904, 2012CB966704, 2012CB966604), the National Natural Science Foundation of China (31425012, 31371185, 81400076, 31322037, 81561138005, 81421002, 81561138003, 81370596, 91439128), National Key Program on Stem Cell and Translational Research (SQ2016ZY05002341), and a SKLEH-Pilot Research Grant (ZK12-04 and ZK13-04).

Author Contributions B.L., F.T., and W.Y. designed the study. F.Z. performed the pre-HSC-related experiments with help from Z.L. and X.Z.; W.W. and W.H. performed the HSC transplantation and HPC assay of Rictor mutant embryos with help from X.M.; Y.N. performed the flow cytometry with help from F.Z.; X.L. performed the single-cell RNA-sequencing; and P.Z. and J.Z. performed the bioinformatics analysis with help from F.X., M.D., and L.W. B.L., F.T., W.Y., and Y.L. wrote the manuscript with support from T.C.

Author Information All of the single-cell and ten-cell RNA-seq data have been deposited in Gene Expression Omnibus under accession numbers GSE67120 and GSE66954. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to W.Y. (wpyuan@ihcams.ac.cn) or F.T. (tangfuchou@pku.edu.cn) or B.L. (bingliu17@yahoo.com).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Mice. Mice experiments were approved by the ethics committee of the affiliated hospital of Academy of Military Medical Sciences. The morning of detection of vaginal plug was defined as E0. The number of somite pairs of every embryo was counted under the microscope during dissection and only the embryos with 41–45 somite pairs were used in E11 pre-HSC experiments. For transplantation assays, male CD45.1/1 mice were crossed with female CD45.2/2 mice to obtain CD45.1/2 embryos. Tie2-Cre and Vav-Cre mice were purchased from Jackson Laboratory. Rictor floxed mice were provided by M. A. Magnuson. Male *Tie2-Cre;Rictor^{fl/fl}* and female *Rictor^{fl/fl}* mice were used to obtain conditional *Rictor^{-/-}* embryos on a CD45.2 background.

Flow cytometry. Cells were analysed and sorted by flow cytometer Calibur or Arial 2 (BD Bioscience). Data were analysed with FlowJo software (Tree Star). The following antibodies were used for staining of the cells: CD31 (MEC13.3), CD41 (MWRReg30), CD45 (30-F11), c-Kit (2B8), CD201 (eBio 1560), CD45.1 (A20), CD45.2 (104), CD3 (145-2C12), Gr-1 (RB6-8C5), Mac-1 (M1/70), B220 (RA3-6B2), CD4 (GK1.5), CD8 (53-6.7), CD47 (miap301), AA4.1 (AA4.1), Ter119 (TER-119), VE-cadherin (eBioBV13), Sca-1 (D7), CD127 (A7R34), CD4 (GK1.5), CD48 (HM48-1), CD8a (53-6.7), Gr-1 (RB6-8C5), CD150 (TC15-12F12.2), CD168 (Proteintech), Ki67 (SolA15), 7-amino-actinomycin D (7-AAD) and Streptavidin APC-eFluor 780. All monoclonal antibodies and 7-AAD were purchased from eBioscience, except for CD31 and CD41 from BD Pharmingen, and CD150 and VE-cadherin from BioLegend. For single-cell sorting, rainbow beads were used to confirm that exactly one single cell was sorted into each well. For cell-cycle-related sorting, Hoechst 33342 (Sigma) and Pyronin Y (Sigma) staining was performed at 37 °C before surface marker staining and BD influx sorter were used.

OP9-DL1 or OP9 co-culture. The OP9-DL1 and OP9 stromal cell line were a gift from Y. Zhang. The OP9-DL1 cells were cultured in a 24-well plate 1 day before the FACS-sorted cells were seeded. The media for co-culture was composed of a-MEM (Gibco), 10% fetal bovine serum (Hyclone), and cytokines (100 ng ml⁻¹ SCF, 100 ng ml⁻¹ IL-3 and 100 ng ml⁻¹ Flt3 ligand, all from PeproTech). After 6 days of culture, the cells in each well were harvested separately for further transplantation. For OP9 co-culture, sorted ECs were plated on an OP9 stromal layer in the presence of 50 ng ml⁻¹ SCF, 10 ng ml⁻¹ IL3, 10 ng ml⁻¹ Flt3 ligand, and 10 ng ml⁻¹ IL-7 (all from PeproTech). After 7 days of culture, co-cultured cells were harvested for flow cytometry analysis.

HSC transplantation. Mouse AGM region dissection and preparation for single-cell suspensions were described elsewhere⁴. Eight- to ten-week-old female C57BL/6 (CD45.2/2) mice were exposed to a split dose of 9 Gy γ -irradiation (⁶⁰Co). The co-cultured cells with OP9-DL1 or freshly isolated cells (CD45.1/2), together with 2 × 10⁴ nucleated bone marrow cells (CD45.2/2), were injected into irradiated adult recipients (CD45.2/2) via the tail vein. The recipients demonstrating ≥5% donor-derived chimaerism in peripheral blood were counted as successfully reconstituted.

Preparation of single-cell RNA-seq library. FACS-enriched cells were kept on ice until lysed and reverse transcribed. Morphologically deformed cells were discarded. Single cells were rinsed in PBS-BSA, and manually transferred into cell lysis buffer with a mouth pipette; 0.05 µl of 1:200,000 dilution of External RNA Controls Consortium (ERCC) RNA spike-in Mix1 (Ambion) was added to lysis buffer per reaction. The cDNA libraries from single cells or ten-cell pools were generated as described previously⁵⁰. Fifty to 200 ng of amplified single-cell or ten-cell-pool cDNAs were sonicated to ~250-base-pair (bp) fragments by the Covaris S2 system, and libraries were generated using a NEBNext Ultra DNA Prep Kit for Illumina (New England Biolabs, E7370) following the manufacturer's protocol.

Single-cell RNA-seq data analysis. All RNA-seq raw data were first pre-processed to remove Illumina adaptor sequences, amplification primer (UP1: ATATGGATCCGGCGCGCCGTCGACTTTTTTTTTTTTTTTTTTTTTTTT; UP2: ATATCTCGAGGGCGCGCCGATCCTTTTTTTTTTTTTTTTTTTT) and 3' polyA sequences using cutadapt version 1.6 and followed by sequence alignment to the mouse genome (mm9) and spiked-in sequences using Tophat version 2.0.12 with the default parameters^{51,52}. Cufflinks version 2.2.1 was used to estimate the expression level of each detected gene as FPKM value⁵³.

To estimate the limit of detection of expression level and quantified copy number of each transcript in single cells, we fitted a linear model for each single cell using the prior known concentration of ERCC spike-in transcripts and FPKM values detected using RNA-seq. The quantified RNA copy number of each expressed gene was calculated according to the FPKM value and the fitted linear formula. To ensure the accuracy of estimated FPKM values and remove the auxiliary data,

only genes with FPKM > 1 in at least one sample were analysed. Expression levels of each gene in all samples were log₂ converted in the following analysis.

Unsupervised hierarchical clustering analysis of single cells was done to evaluate the accuracy of FACS results and detected outliers were removed. PCA and further determined markers of different cell groups also supported the clustering results. In total, 99 single cells were retained for ECs with 17 individual cells, T1 pre-HSCs with 26 cells, T2 pre-HSCs with 19 cells, E12 HSCs with 21 cells, and E14 HSCs with 16 cells, respectively.

Multiple *t*-tests were used to evaluate the statistical significance of differentially expressed genes for different cell groups. Corrected *P* values (FDR) were calculated using the Benjamini and Hochberg method. Differentially expressed genes were defined only if the corresponding *P* values were less than 0.05 and the FDR was less than 0.05 with a fold change of log₂-converted FPKM larger than 2. GO analysis of differentially expressed genes was performed using DAVID⁵⁴.

Ideally, a cell-type specifically expressed gene is highly and uniformly expressed (with relative expression level as 1) in all cells in a particular cell type (for example T1 pre-HSCs) but not expressed (with relative expression level as 0) in all cells of other cell types (for example ECs, T2 pre-HSCs, and E12 and E14 HSCs). In reality, a cell-type specifically expressed gene will be the one that is highly expressed in most cells in a particular cell type (while lowly or not expressed in the remainder of cells of this type), and lowly or not expressed in all cells of other cell types. These real 'cell-type specifically expressed genes' will show high correlation with an ideal 'cell-type specifically expressed gene'.

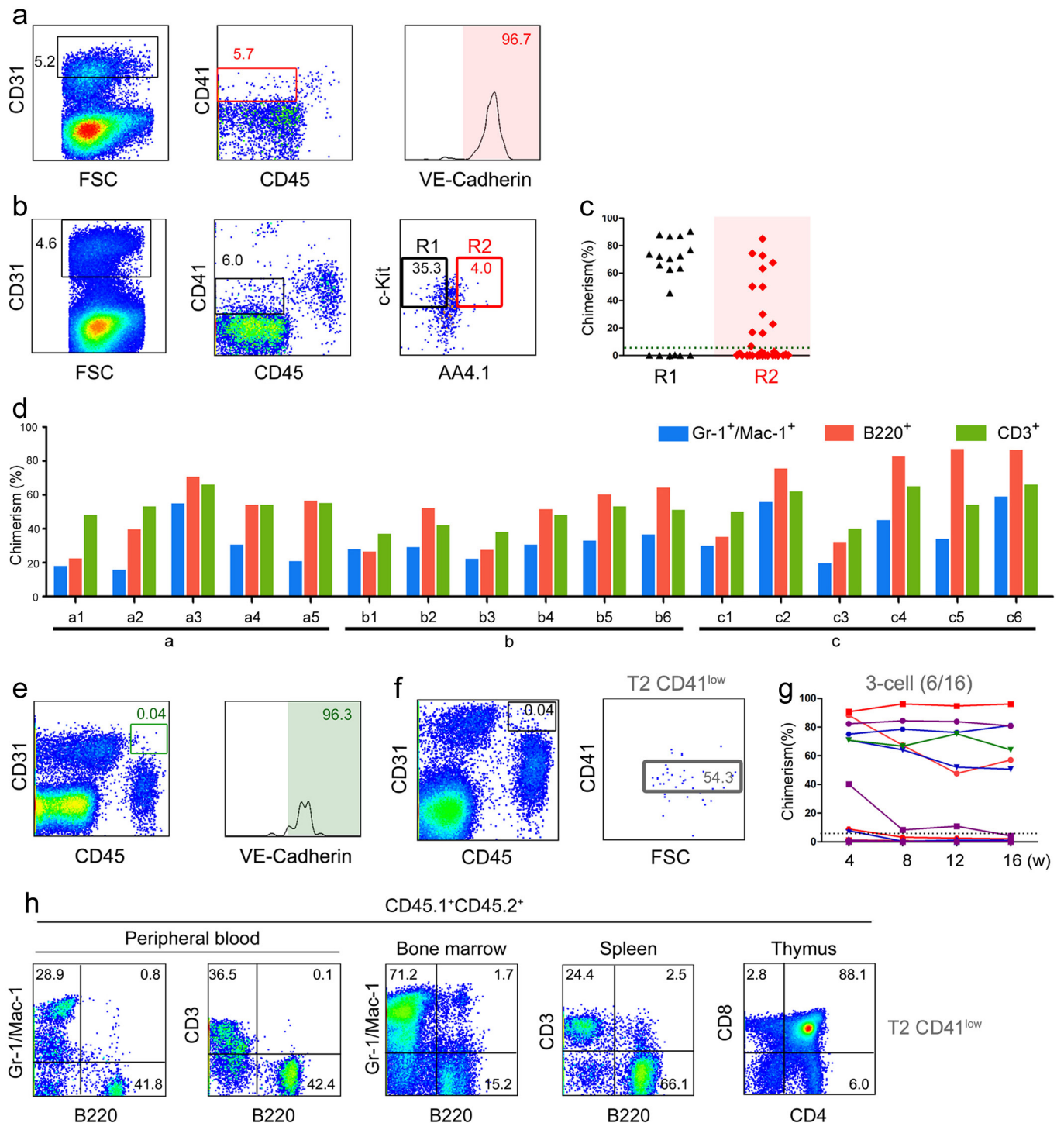
So, for each cell type, we artificially defined an ideal gene with its expression pattern as follows: with relatively high expression (relative expression level as 1) in all cells of this cell type (for example T1 pre-HSCs) and with no expression (relative expression level as 0) in all cells of other cell types (for example ECs, T2 pre-HSCs, and E12 and E14 HSCs). Then all of the real genes were defined as cell-type specifically expressed ones if the Pearson correlation coefficients between those genes and this 'ideal' cell-type specifically expressed gene were higher than 0.5.

To find genes that contributed most in separating different cell groups, we took advantage of PCA. We only considered genes with a coefficient of variance larger than 0.5. PCA was applied to all remnant genes, and the top 300 genes that contributed most in both principal components 1 and 2 were used to explore the expression patterns by the hierarchical clustering method. After the expression patterns were defined, the median expression value of the most contributing genes that clustered together was calculated as the representative expression level of the corresponding pattern in each cell. Pearson's correlation coefficient was calculated for all genes to each pattern. Thus, genes that contributed most to separate different cell groups were determined. Pattern-related transcription factors were also determined using this strategy. We next calculated the network adjacency of transcription factors and all expression patterns using WGCNA⁵⁵ with a threshold 0.02 to include edges for visualization display in Cytoscape⁵⁶.

The *t*-distributed stochastic neighbour embedding was used to display high-dimensional gene expression profiles of 99 single cells. The R package 'tsne' was applied after calculating the 'Manhattan' distance of each pair of samples. To order individual cells through HSC development by their dynamic transcriptomes instead of predefined cell type, pseudo-time analysis was performed using Monocle as in a previous study²⁷.

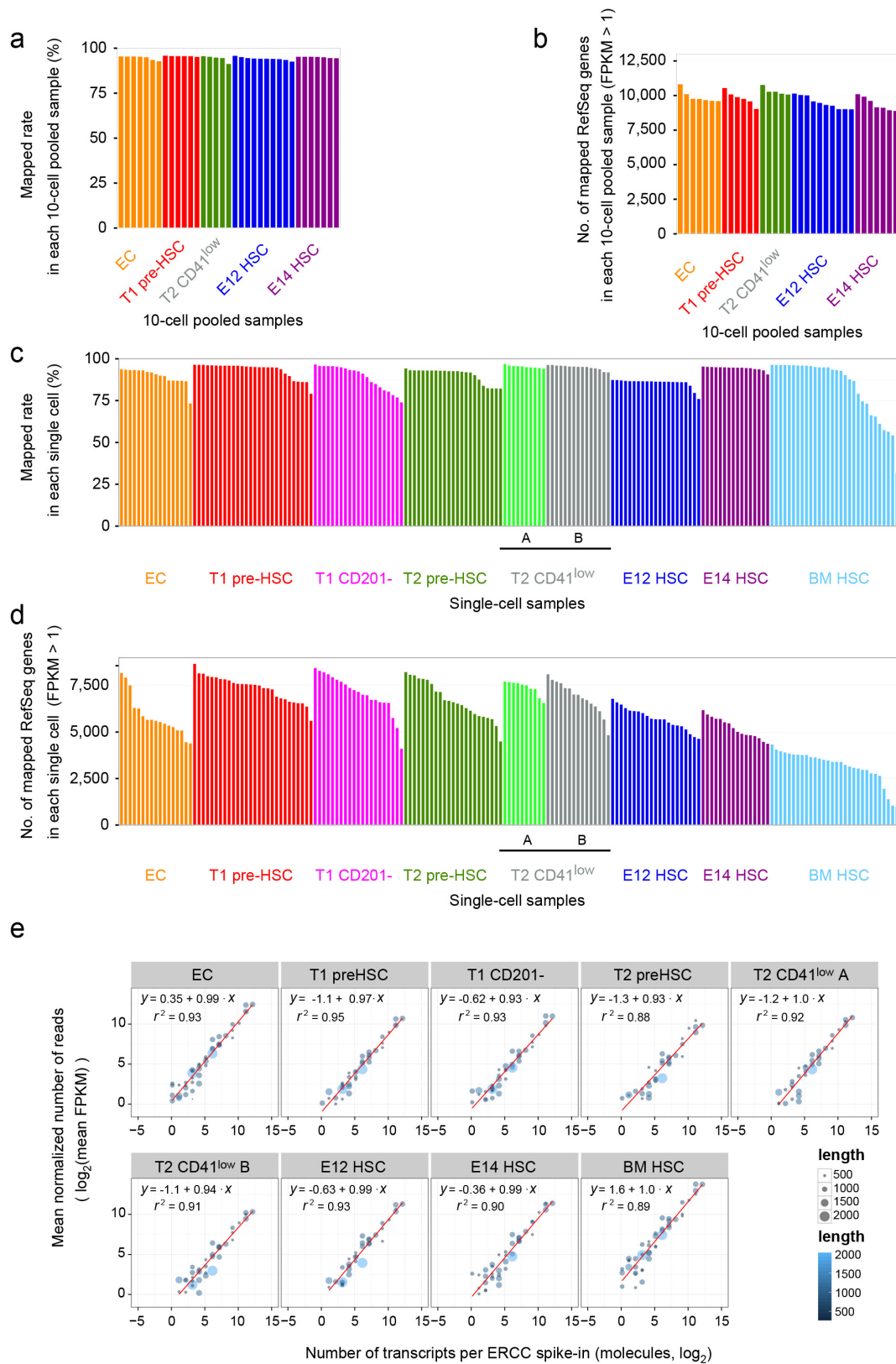
GSEA/KEGG analysis. GSEA is a computational pathway analysis tool that determines whether a set of genes show statistically significant, concordant differences between two biological states (<http://www.broadinstitute.org/gsea/index.jsp>). To create gene sets for a genome with custom annotations, we associated our genes with known KEGG pathways and made each pathway a gene set (<http://www.genome.jp/kegg/pathway.html>), then used the 'Signal2Noise' ranking metric and chose the gene sets showing significant change at FDR < 0.25 and nominal *P* value < 0.05.

50. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**, 377–382 (2009).
51. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
52. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
53. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnol.* **28**, 511–515 (2010).
54. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**, 44–57 (2008).
55. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
56. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).



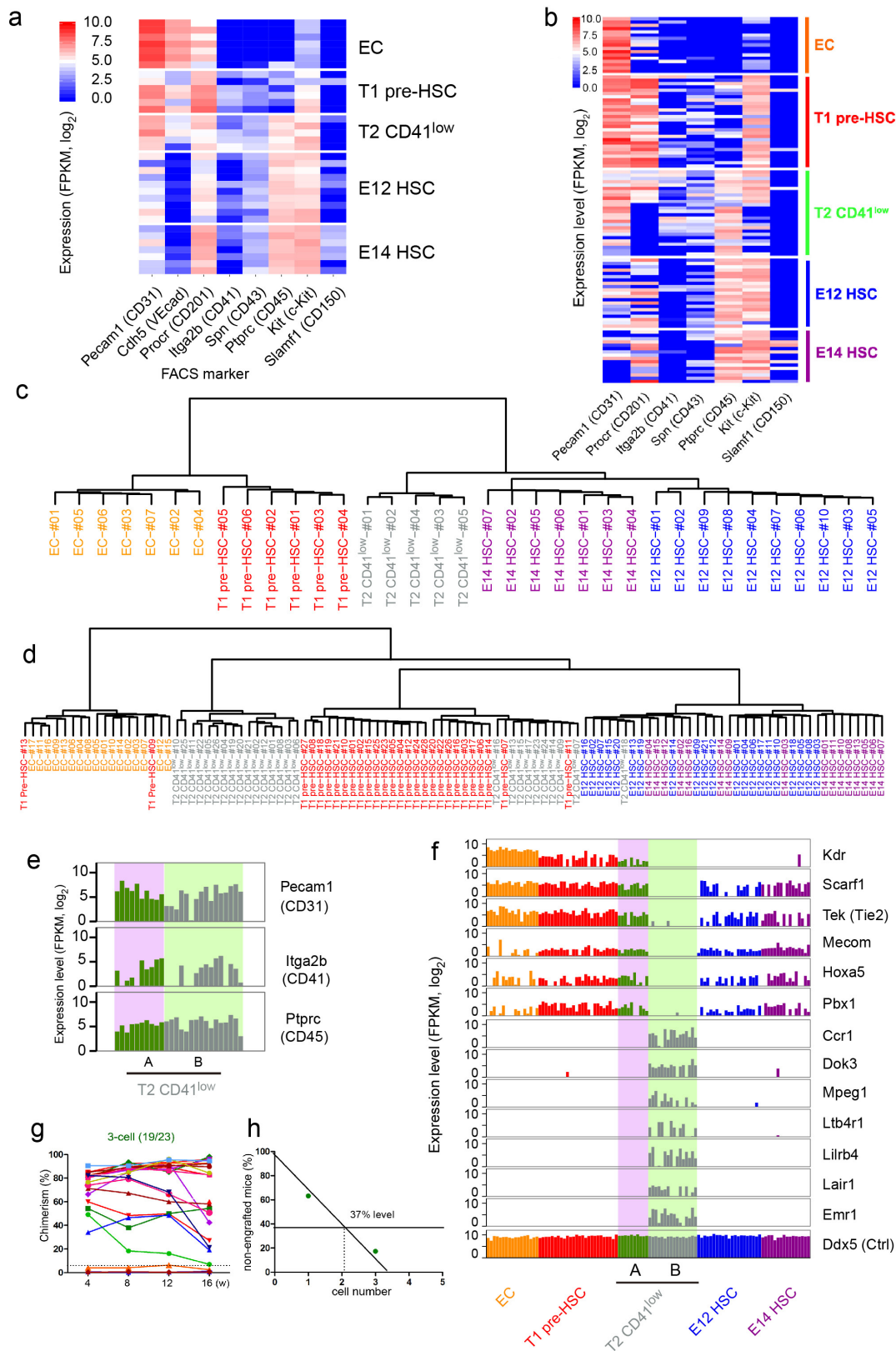
Extended Data Figure 1 | Identification of pre-HSCs in the E11 AGM region. **a**, Expression of VE-cadherin in CD31⁺CD41^{low}CD45⁻ cells of the E11 AGM region. **b**, FACS isolation of T1 pre-HSCs on the basis of AA4.1 expression. **c**, Donor chimerism in peripheral blood of recipients receiving cultures of CD31⁺CD45⁺CD41^{low}c-Kit⁺AA4.1⁻ and CD31⁺CD45⁺CD41^{low}c-Kit⁺AA4.1⁺ cells, respectively, 4 weeks after transplantation. **d**, Multi-lineage repopulation (8–12 weeks) in peripheral blood of secondary recipients ($n = 17$) receiving HSCs from three reconstituted primary recipients by single T1 pre-HSC-derived co-cultures. Donor-derived (CD45.1⁺CD45.2⁺) chimerism in myeloid (Gr-1⁺/Mac-1⁺), B-lymphoid (B220⁺), and T-lymphoid (CD3⁺) cells are shown, and 'a', 'b' and 'c' indicate three different mice reconstituted by

single-cell co-cultures. **e**, Expression of VE-cadherin in CD31⁺CD45⁺ cells of the E11 AGM region. **f**, FACS isolation of E11 T2 pre-HSCs with CD41. **g**, Donor chimerism in peripheral blood of recipients receiving 6-day co-cultures from CD31⁺CD45⁺CD41^{low} population (three cells per recipient, $n = 16$) monitored at 4, 8, 12, and 16 weeks after transplantation. **h**, Multi-lineage long-term (>16 weeks) repopulation in peripheral blood, bone marrow, spleen, and thymus of primary recipients transplanted with 6-day co-cultures initiated from the CD31⁺CD45⁺CD41^{low} population. Donor-derived (CD45.1⁺CD45.2⁺) myeloid (Mac-1⁺/Gr-1⁺), B-lymphoid (B220⁺), and T-lymphoid (CD3⁺ or CD4⁺/CD8⁺) cells are shown in major haematopoietic organs of a representative recipient.



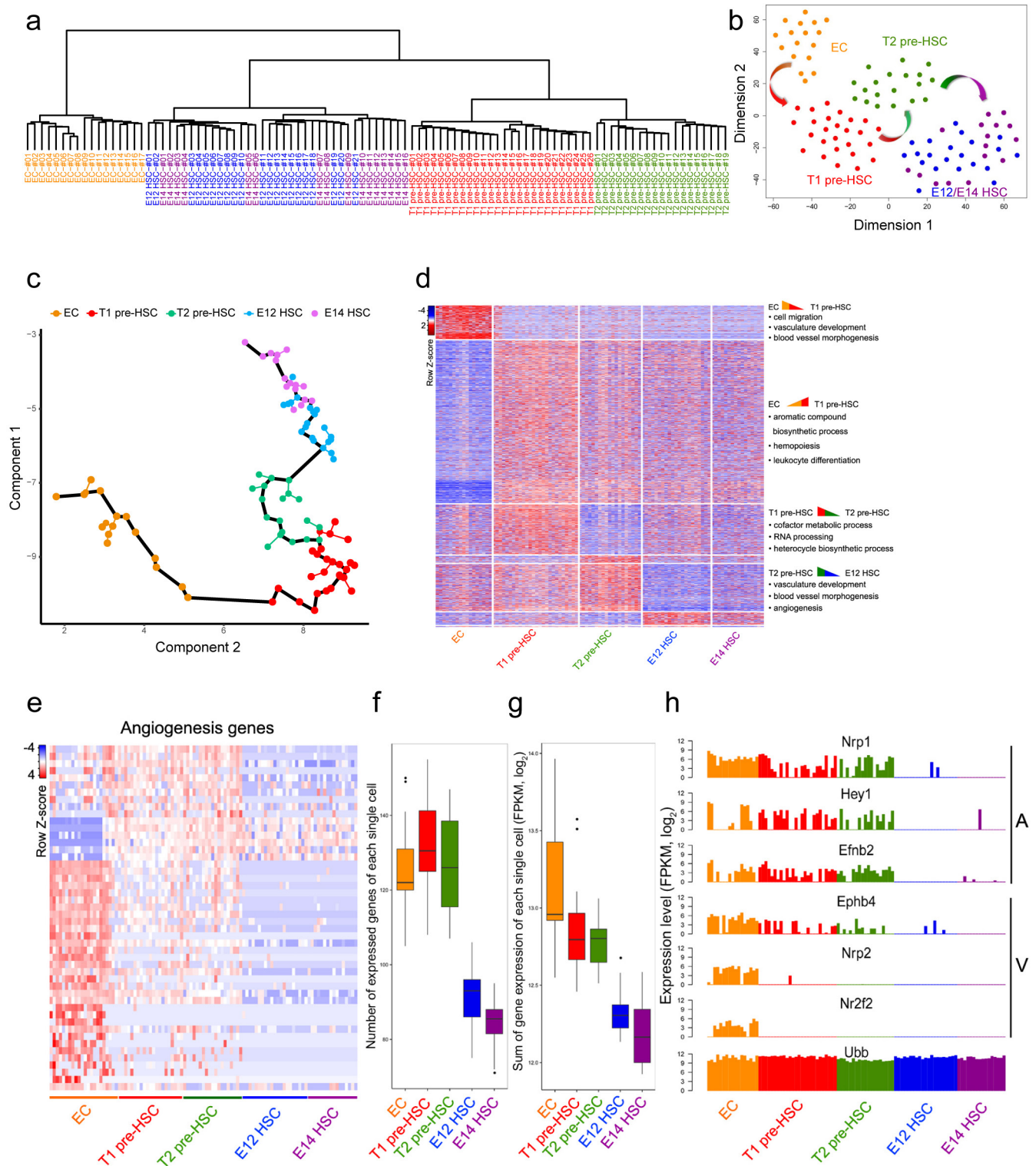
Extended Data Figure 2 | Quality control of the RNA-seq data of single-cell and ten-cell samples. **a**, Bar plot of mapping rate to mouse mm9 reference genome of each ten-cell sample. **b**, Number of detected RefSeq genes (FPKM > 1) in each ten-cell sample. **c**, Bar plot of mapping rate to mm9 reference genome of each single-cell sample. **d**, Number of detected

RefSeq genes (FPKM > 1) of each single-cell sample. **e**, Regression fits between average expression level ($\log_2(\text{mean FPKM})$) and logarithm-transformed copy number of ERCC RNA molecules spiked into lysis of each single-cell sample. All detected ERCC spike-ins above the expression threshold (FPKM > 1) were used for analysis.



Extended Data Figure 3 | Gene expression dynamics during HSC formation. **a**, Heat map of expression levels of the FACS markers for ten-cell-pool RNA-seq analysis. **b**, Heat map of expression levels of FACS markers for single-cell RNA-seq analysis. **c**, Unsupervised hierarchical clustering of transcriptome profiles of 35 ten-cell samples showing that the samples were accurately grouped together according to their cell types. **d**, Unsupervised hierarchical clustering of transcriptome profiles of 108 single-cell samples comprising ECs, T1 pre-HSCs, T2 CD41^{low} cells, E12 HSCs, and E14 HSCs. **e**, Bar plot of RNA expression of Pecam1

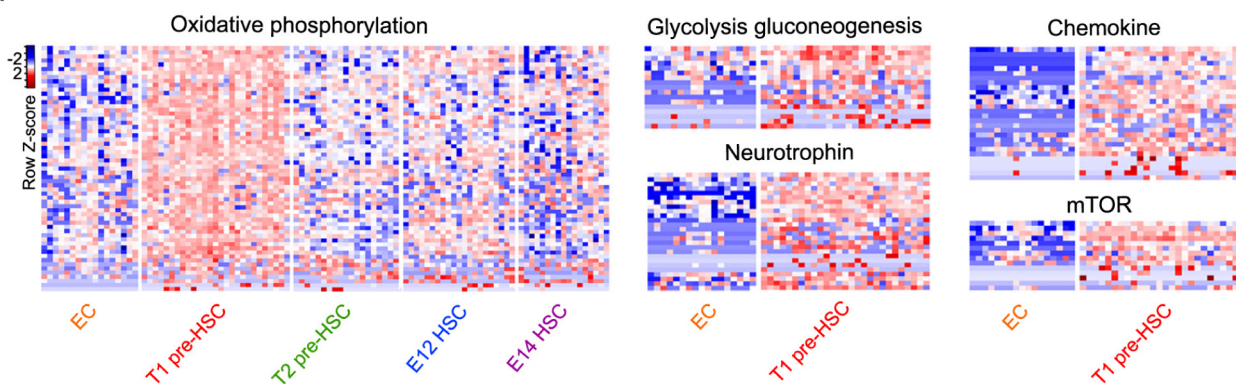
(CD31), Itga2b (CD41), and Ptprc (CD45) in the two subpopulations of T2 CD41^{low} cells. **f**, Endothelial or HSC and HPC markers specifically expressed in CD201-positive T2 CD41^{low} cells as well as other HSC-competent populations, and granulocyte and macrophage markers specifically expressed in the CD201-negative T2 CD41^{low} subpopulation. **g**, Donor-derived chimerism in peripheral blood of recipients in the three-cell T2 pre-HSC group (CD31⁺CD45⁺c-Kit⁺CD201^{high}) monitored at 4, 8, 12, and 16 weeks after transplantation. **h**, Quantification of T2 pre-HSC frequency by limiting dilution analysis.



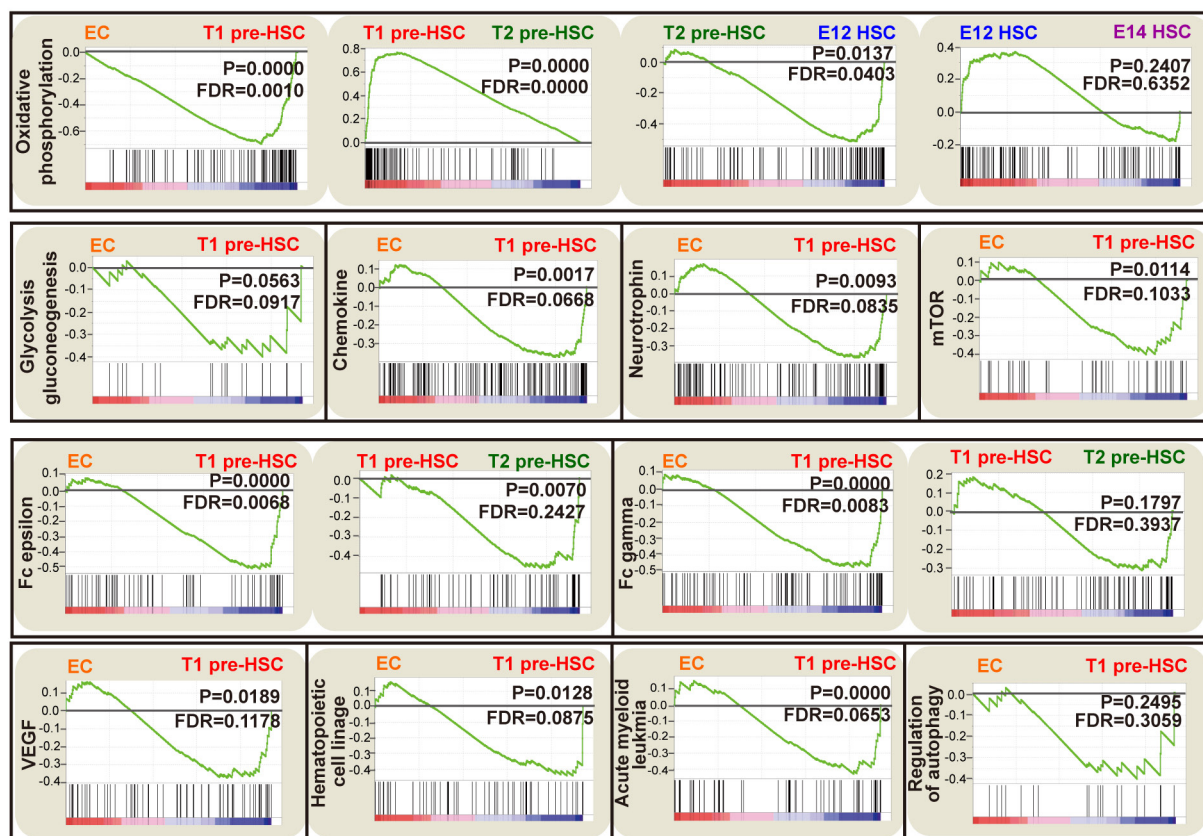
Extended Data Figure 4 | Global and angiogenesis-related gene expression during HSC formation. **a**, Unsupervised hierarchical clustering of 99 single cells. **b**, The *t*-distributed stochastic neighbour embedding display of transcriptome profiles of 99 single cells, indicating three major groups as ECs, pre-HSCs, and mature HSCs. **c**, Pseudo-time analysis of 99 single cells by the Monocle method. **d**, Heat map of differentially expressed genes between each of the two consecutive stages. The major biological process GO terms enriched in differentially

expressed genes are shown to the right. **e**, Heat map of 58 angiogenesis genes with dynamic expression changes in different cell types. **f**, Box plot of the number of expressed angiogenesis genes in each individual cell of different cell types. Genes with expression level FPKM > 1 are defined as expressed genes in a single cell. **g**, Box plot of total expression level (total FPKM) of all of the angiogenesis genes expressed in each individual cell. **h**, Bar plot of the expression of selected artery (A) and vein (V) genes in each single cell.

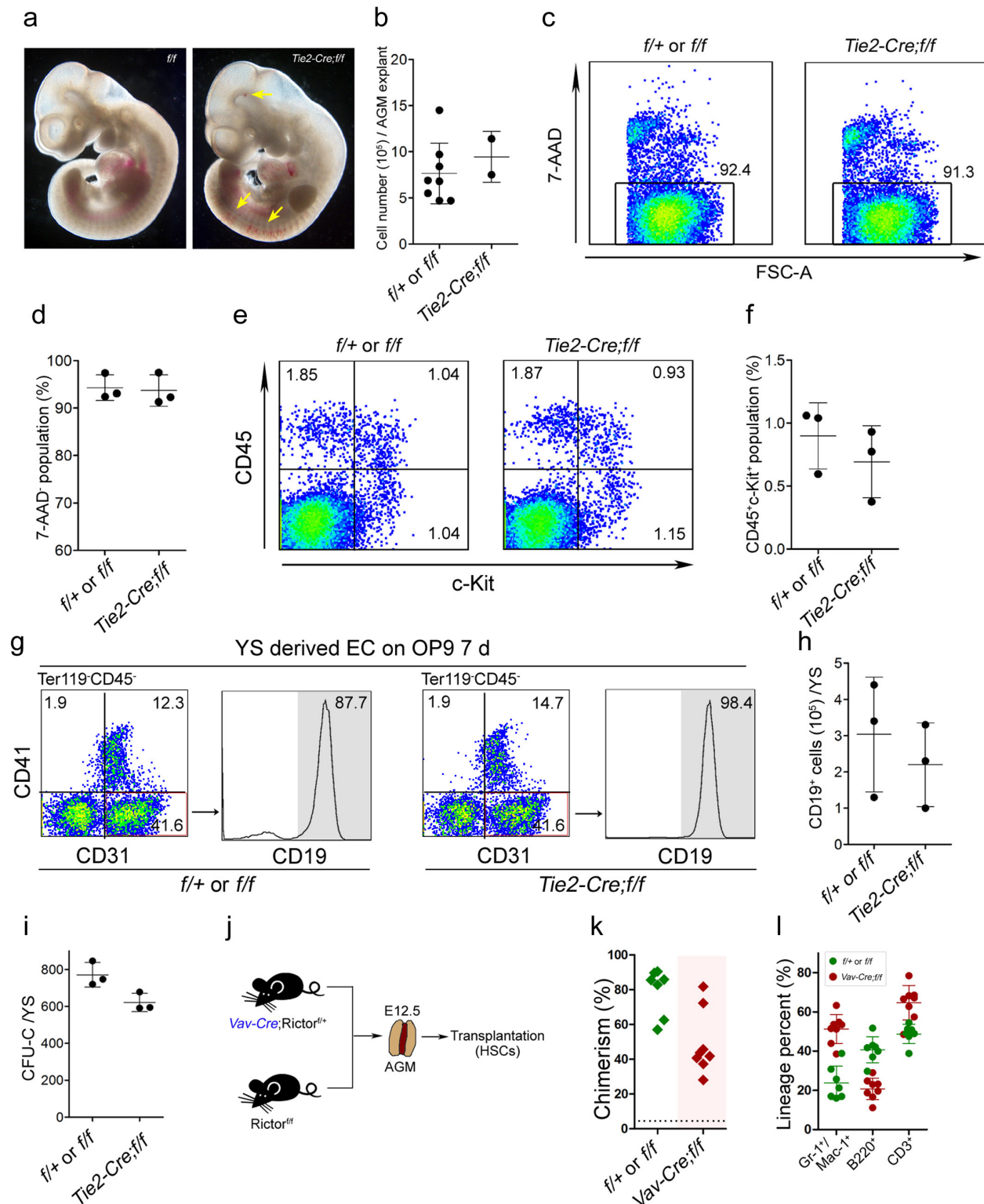
a



b

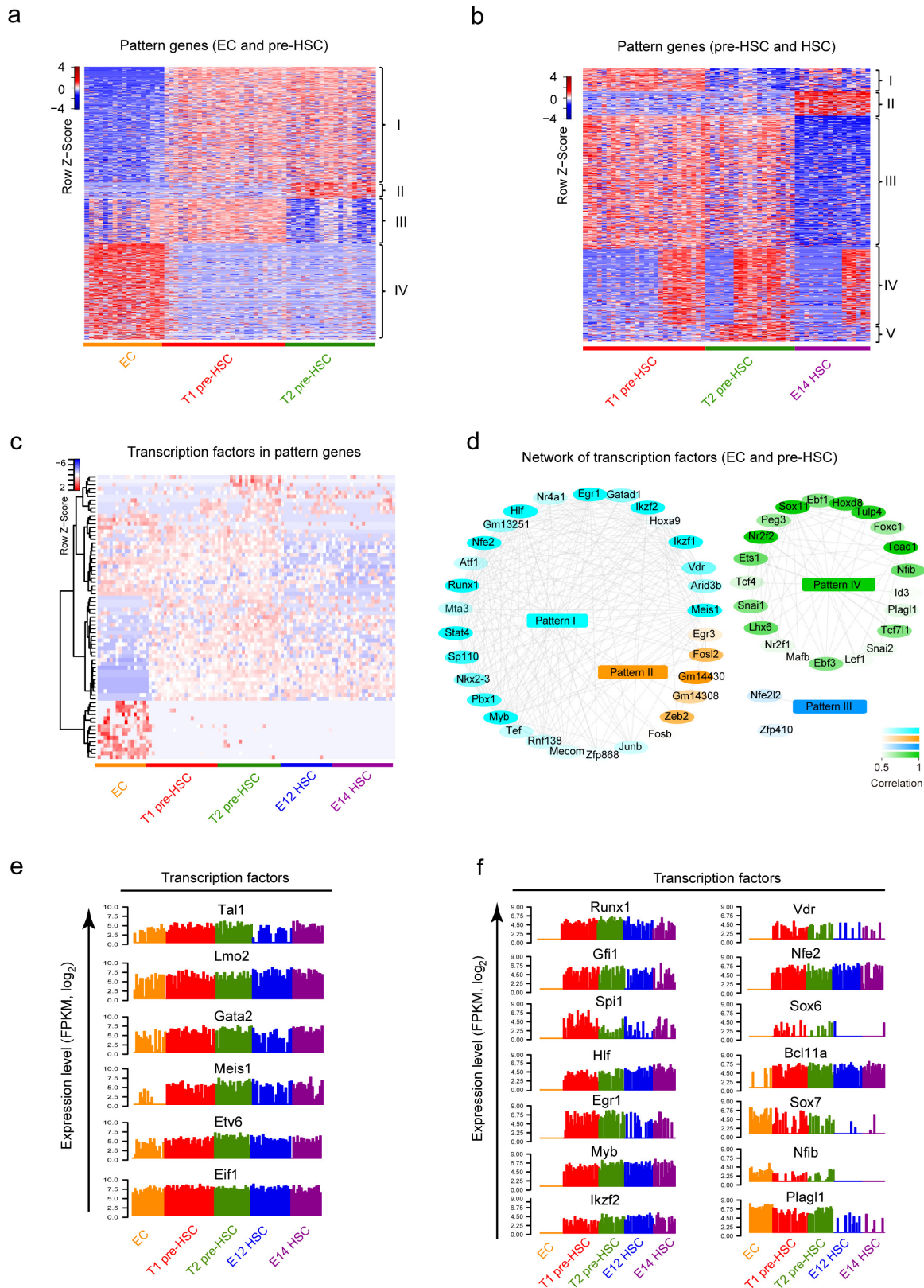


Extended Data Figure 5 | Signalling pathways enriched in pre-HSCs by GSEA/KEGG analysis. a, Heat maps of genes enriched in representative signalling pathways. **b,** GSEA enrichment plot of KEGG signalling pathways. Nominal P value, empirical phenotype-based permutation test ($P < 0.05$, FDR < 0.25).



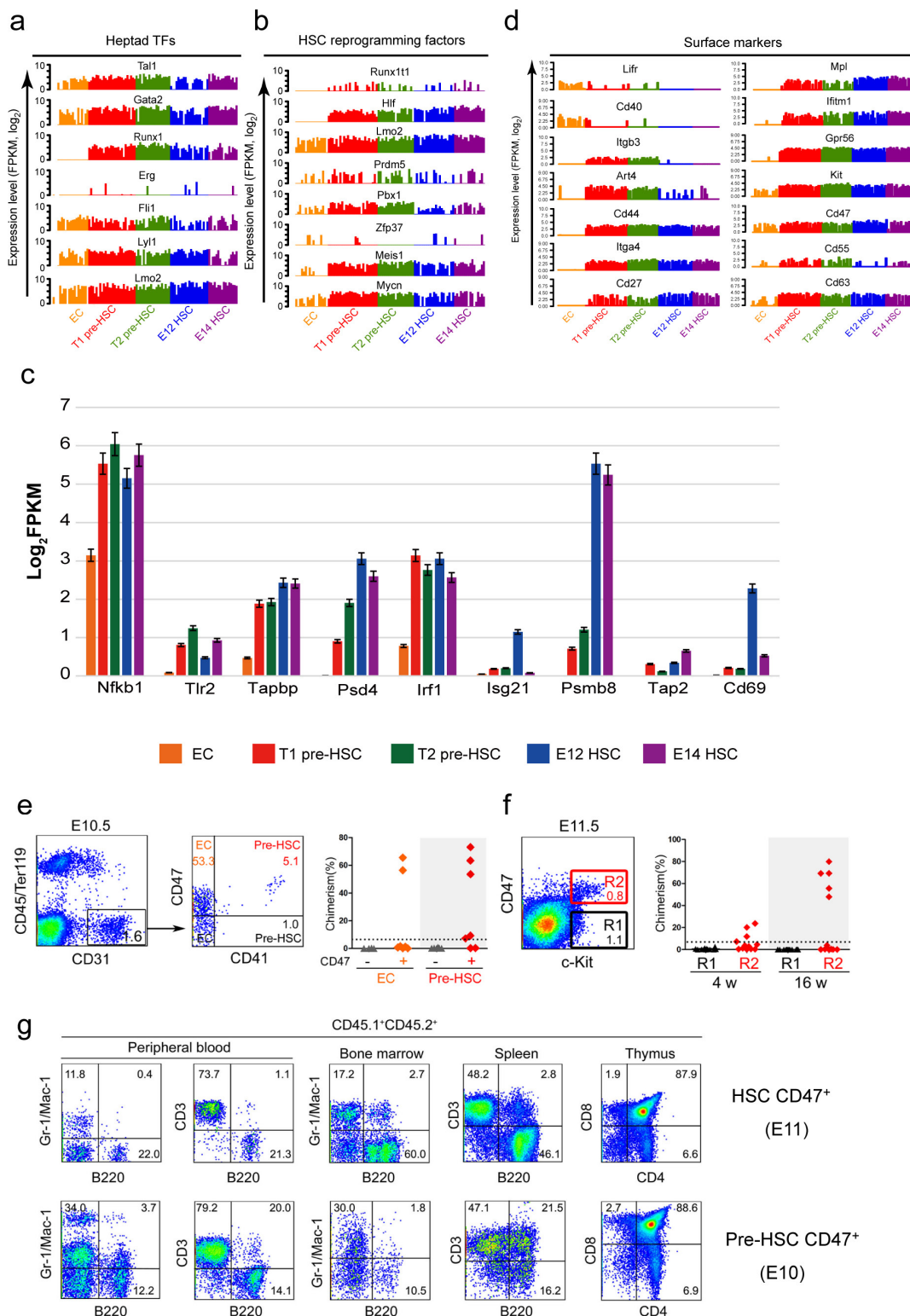
Extended Data Figure 6 | Morphology and haematopoietic potential of *Rictor* mutant embryos. **a**, Morphology of E10 *f/f* (37 somite pairs) and *Tie2-Cre;f/f* (39 somite pairs) embryos. The arrows indicate scattered haemorrhage. **b**, Cell number of E10 AGM after 72-h organ culture. Data are collected from eight (*f/+* or *f/f*) and two (*Tie2-Cre;f/f*) embryos. **c**, **d**, Representative FACS analysis and quantification of 7-AAD⁻ cells in E10 *Tie2-Cre;Rictor^{fl/fl}* and control AGM after 72-h organ culture. Data are collected from three independent experiments using 11 (*f/+* or *f/f*) and 7 (*Tie2-Cre;f/f*) embryos. **e**, **f**, FACS analysis and quantification of CD45⁺ c-Kit⁺ cells in *Tie2-Cre;Rictor^{fl/fl}* and control AGM after organ culture. Data are collected from 3 independent experiments using 11 (*f/+* or *f/f*) and 7 (*Tie2-Cre;f/f*) embryos. **g**, **h**, B-lymphoid potential of the immunophenotypically defined ECs

purified from the E10 yolk sac (YS) after co-culture with OP9 stromal cells. Data are mean \pm s.d. of three independent experiments using six embryos per genotype. **i**, Quantification of CFU-Cs in the E10 yolk sac. Data are collected from three CFU-C cultures per genotype in a representative of two independent experiments. **j**, Schematic experimental design. **k**, Reconstitution potential of E12.5 AGM in *Vav-Cre;Rictor^{fl/fl}* relative to the controls. Symbols represent the donor chimaerism of CD45.2⁺ cells in peripheral blood of individual recipients 4 months after transplantation. Data are collected from seven (*f/+* or *f/f*) and eight (*Vav-Cre;f/f*) recipients. **l**, Donor chimaerism of myeloid (Gr-1⁺/Mac-1⁺), B-lymphoid (B220⁺) and T-lymphoid (CD3⁺) cells repopulated by the E12.5 AGM region of *Vav-Cre;Rictor^{fl/fl}* and control embryos after 4 months of transplantation.



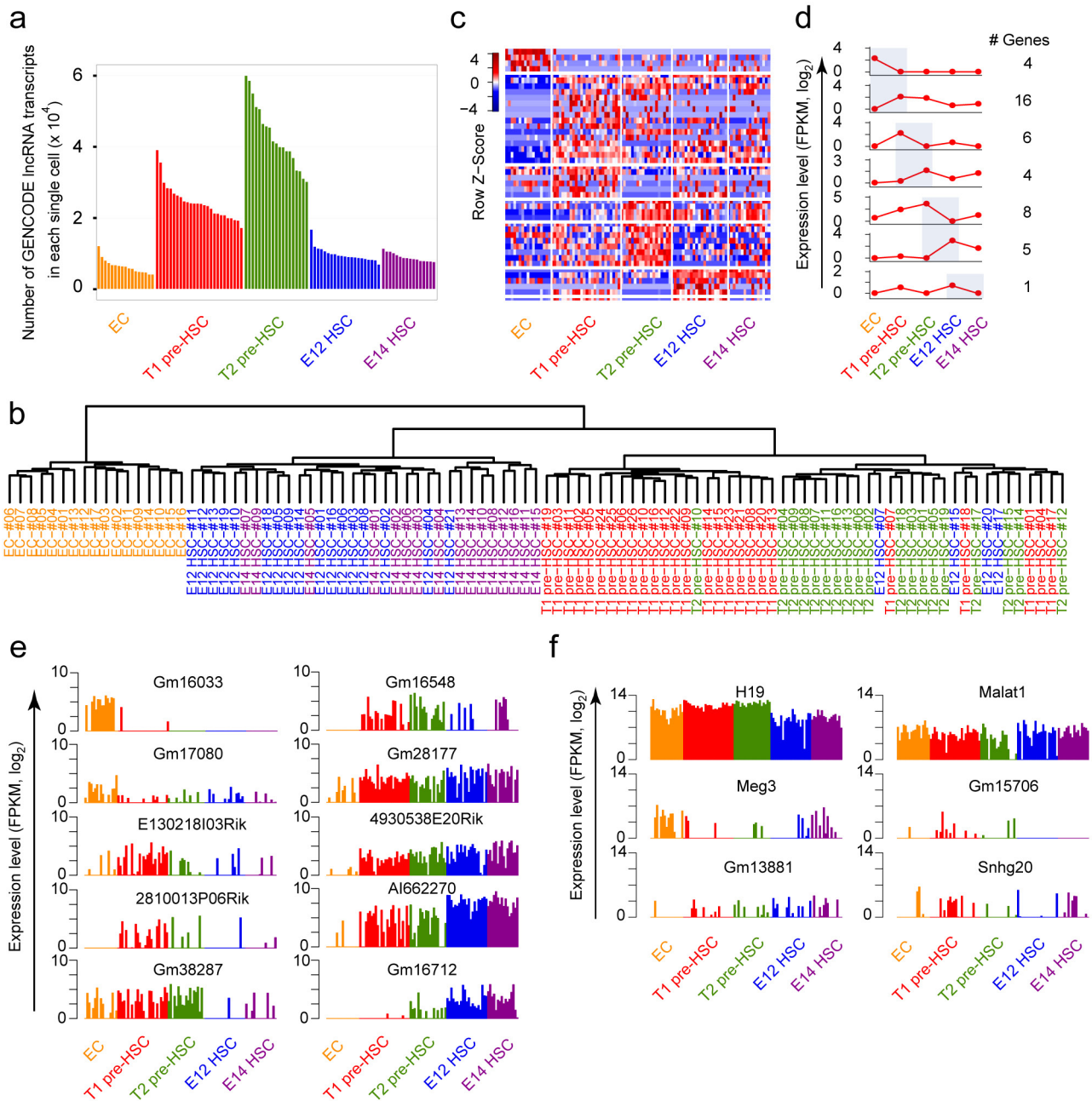
Extended Data Figure 7 | Dynamic expression patterns and network of transcription factors during HSC formation. **a**, Heat map of genes showing different expression patterns in ECs and pre-HSCs. Note four distinct expression patterns as highly expressed in pre-HSCs (I), highly expressed in T2 pre-HSCs (II), highly expressed in T1 pre-HSCs (III), and highly expressed in ECs (IV). **b**, Heat map of genes showing different expression patterns in pre-HSCs (T1 and T2) and mature HSCs (E12 and E14). Note five distinct expression patterns as highly expressed in T1 pre-HSCs (I), highly expressed in E14 HSCs (II), highly expressed in pre-HSCs (III), and

heterogeneous in all cell types (IV), and highly expressed in T2 pre-HSCs (V). **c**, Heat map of expression dynamics of transcription factors highly related to the above patterns in these single cells. **d**, Network view of transcription factors (TFs) related to different expression patterns determined in Fig. 5b. Transcription factors were arranged using circle layout in Cytoscape. A deeper background colour of the gene names indicates higher correlations of transcription factors to that expression pattern. **e**, **f**, Bar plot of expression dynamics of selected transcription factors in single cells showing different expression patterns in distinct cell types.



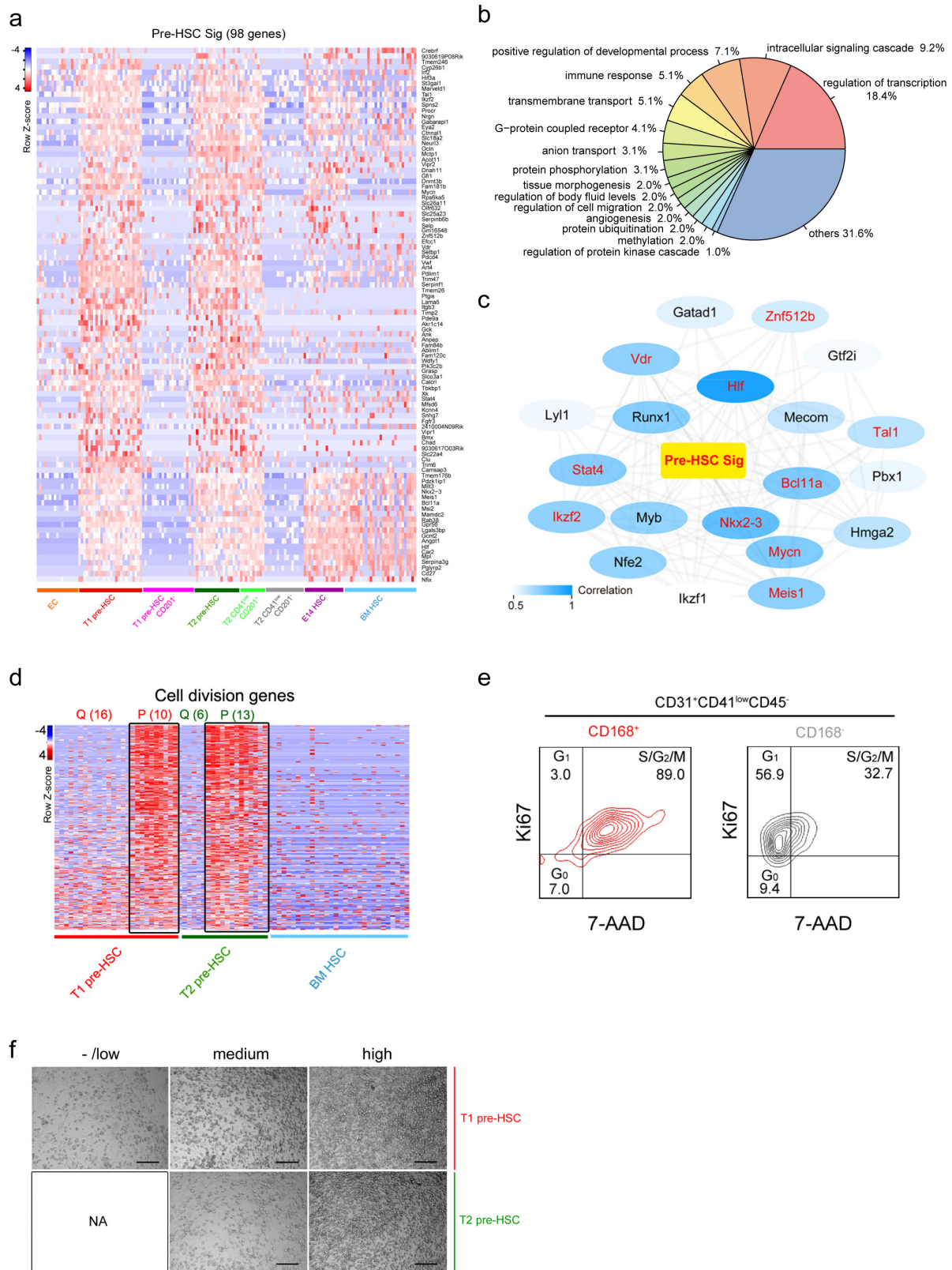
Extended Data Figure 8 | Expression of functional molecules and surface markers during HSC formation. **a**, Bar plot of expression of the documented heptad transcription factors in each single cell during HSC formation. **b**, Bar plot of expression of the definitive HSC-reprogramming factors in each single cell during HSC formation. **c**, Mean value of expression (FPKM, log₂) of upregulated innate immune/inflammatory genes in pre-HSCs and/or HSCs. **d**, Bar plot of expression dynamics of selected surface markers in single cells showing different expression patterns in distinct cell types. **e**, FACS sorting of E10 mouse AGM cells into four subpopulations: CD47⁺ ECs, CD47⁻ ECs, CD47⁺ pre-HSCs (CD47⁺CD31⁺CD41⁺CD45⁻Ter119⁻), and CD47⁻ pre-HSCs (CD47⁻CD31⁺CD41⁺CD45⁻Ter119⁻). Symbols represent the donor chimerism in peripheral

blood at 16 weeks after transplantation of co-cultured cells with the four subpopulations, respectively. **f**, FACS sorting of c-Kit⁺CD47⁺ and c-Kit⁺CD47⁻ subpopulations from the E11 mouse AGM region. Symbols represent the donor chimerism in peripheral blood of recipients at 4 and 16 weeks after direct transplantation of c-Kit⁺CD47⁺ and c-Kit⁺CD47⁻ subpopulations. **g**, Multi-lineage long-term (>16 weeks) repopulation in peripheral blood, bone marrow, spleen, and thymus of recipients transplanted with E11 HSCs (c-Kit⁺CD47⁺), or with 6-day co-cultures initiated from E10 pre-HSCs (CD31⁺CD45⁻Ter119⁻CD41⁺CD47⁺). Donor-derived (CD45.1⁺CD45.2⁺) myeloid (Mac-1⁺/Gr-1⁺), B-lymphoid (B220⁺), and T-lymphoid (CD3⁺ or CD4⁺/CD8⁺) cells are shown in major haematopoietic organs of a representative recipient.



Extended Data Figure 9 | Expression pattern of lncRNAs during HSC formation. **a**, Bar plot of total copy numbers of the lncRNAs in each individual cell. **b**, Unsupervised hierarchical clustering of lncRNA expression profiles of 99 samples indicating three major groups as ECs, pre-HSCs, and mature HSCs. **c**, Heat map of expression dynamics of 35 differentially expressed lncRNAs of each two consecutive developmental stages. Z score with colour from blue to red indicates expression level from low to high. **d**, Dynamic expression changes of differentially expressed lncRNAs through five developmental stages for each two developmentally

consecutive cell types in shaded areas, with the number of differentially expressed lncRNAs listed on the right panel. **e**, Bar plot of expression dynamics of representative differentially expressed lncRNAs in single cells during HSC formation. **f**, Bar plot of expression pattern of 26 adult HSC-specific lncRNAs which were documented in ref. 40 in single cells during HSC formation. H19 and Malat1 were significantly expressed in most HSC-related cells. The expression of Meg3 was relatively abundant in ECs and E14 HSCs, and higher expression of Gm15706 was observed in the T1 pre-HSCs.



Extended Data Figure 10 | Signature genes and cell-cycle heterogeneity of pre-HSCs. **a**, Heat map of 98 pre-HSC signature genes in various cell types. **b**, Distribution of biological process GO terms of the annotated pre-HSC signature genes. **c**, Network view of transcription factors in pre-HSC signature genes. A deeper background colour of the gene names indicates higher correlations of transcription factors to the signature gene expression pattern. Highlights in red font represent transcription factors in 98 pre-HSC signature genes. **d**, Heat map representation of different cell-cycle status in pre-HSCs and adult HSCs. Q, quiescent; P, proliferative. Numbers within brackets

indicate the number of single-cell RNA-seq samples. **e**, Cell-cycle status analysis of the CD31⁺CD41^{low}CD45⁻ population by Ki67/7-AAD staining. **f**, Distinct proliferation capacities of single T1 (CD31⁺CD45⁻CD41^{low}-c-Kit⁺CD201^{high}) and T2 (CD31⁺CD45⁺-c-Kit⁺CD201^{high}) pre-HSCs. The single-cell co-cultures show differential proliferation capacities *in vitro*, from -/low, medium (hundreds of progeny cells), to high (>1,000 progeny cells) degree. Of note, 11 of the 12 repopulated recipients (four T1 and seven T2 pre-HSCs) received the single-cell co-cultures with high proliferation capacity, and the remaining one (T1 pre-HSC) with medium proliferation capacity.

Carcinoma–astrocyte gap junctions promote brain metastasis by cGAMP transfer

Qing Chen^{1*†}, Adrienne Boire^{1,2*}, Xin Jin^{1†}, Manuel Valiente^{1†}, Ekrem Emrah Er¹, Alejandro Lopez-Soto^{1†}, Leni S. Jacob^{1†}, Ruzeen Patwa¹, Hardik Shah³, Ke Xu⁴, Justin R. Cross³ & Joan Massagué¹

Brain metastasis represents a substantial source of morbidity and mortality in various cancers, and is characterized by high resistance to chemotherapy. Here we define the role of the most abundant cell type in the brain, the astrocyte, in promoting brain metastasis. We show that human and mouse breast and lung cancer cells express protocadherin 7 (PCDH7), which promotes the assembly of carcinoma–astrocyte gap junctions composed of connexin 43 (Cx43). Once engaged with the astrocyte gap–junctional network, brain metastatic cancer cells use these channels to transfer the second messenger cGAMP to astrocytes, activating the STING pathway and production of inflammatory cytokines such as interferon- α (IFN α) and tumour necrosis factor (TNF). As paracrine signals, these factors activate the STAT1 and NF- κ B pathways in brain metastatic cells, thereby supporting tumour growth and chemoresistance. The orally bioavailable modulators of gap junctions meclofenamate and tonabersat break this paracrine loop, and we provide proof-of-principle that these drugs could be used to treat established brain metastasis.

Brain metastases occur in 20–40% of advanced stage cancers and represent the most prevalent adult intracranial malignancy¹. Current clinical management of brain metastases affords limited disease control and most patients succumb to tumour progression less than 12 months after diagnosis^{1,2}; better therapeutic strategies are urgently needed. Recent work has begun to describe the cellular and molecular interactions responsible for brain metastasis. Circulating cancer cells first traverse the blood–brain barrier (BBB)^{3,4} to enter the parenchyma, where they co-opt the microvasculature^{5,6}. However, the vast majority of cancer cells that infiltrate the brain perish, rejected by astrocytes⁶. The astrocyte network serves a protective role in the central nervous system^{7,8}. In brain metastasis, reactive astrocytes generate the protease plasmin and cytotoxic cytokines. Brain metastatic cells counter this defence with serpin inhibitors of plasminogen activator⁶. Yet, astrocyte–cancer cell interactions may not be uniformly antagonistic: brain metastases contain abundant reactive astrocytes⁸, and astrocytes can exert a beneficial effect on cancer cell co-cultures⁹.

Here we show that brain metastatic cells selectively establish Cx43 gap junctions with astrocytes by means of PCDH7. These channels allow for passage of cGAMP from cancer cells to astrocytes to activate STING, an innate immune response pathway to cytosolic double-stranded DNA (dsDNA)¹⁰. The resulting astrocyte production of IFN α and TNF supports growth and chemoresistance in brain metastatic cells. Pharmacological inhibition of these gap junctions in mice suppresses brain metastasis.

Brain metastasis linked to Cx43 gap junctions

Glial fibrillary acidic protein (GFAP)-positive reactive astrocytes are a hallmark of brain metastasis (Fig. 1a). Astrocytes interact in a gap-junction

network with Cx43, one of the principal gap junction proteins in these cells¹¹. Cx43 is present in brain metastases, including cancer cell–astrocyte interfaces (Fig. 1a). In triple-negative breast cancer (TNBC) and non-small cell lung cancer (NSCLC), we found a higher level of Cx43 staining in brain metastases than in primary tumours or normal tissues (Fig. 1b, Extended Data Fig. 1a). To characterize these cancer cell–astrocyte interactions, we used five brain metastatic models derived from mammary (MDA231-BrM2 and ErbB2-BrM) or lung (H2030-BrM3, 393N1 and LLC-BrM) adenocarcinomas, of human or mouse origin^{3,6,12,13} (Extended Data Fig. 1b). These lesions display Cx43 expression at the cancer cell–astrocyte interface (Fig. 1c). In each of these models, co-culture with astrocytes protected cancer cells from chemotherapy and the pro-apoptotic cytokine soluble Fas ligand (sFasL) (Extended Data Fig. 1c), consistent with previous *in vitro* findings⁹ and suggesting a dual role for astrocytes in brain metastasis.

Gap junctions are formed by hexameric connexin hemi-channels and allow for the passage of cytoplasmic molecules between cells¹⁴. We observed time-dependent transfer of calcein from brain metastatic cells to astrocytes by time-lapse fluorescence microscopy (Fig. 1d; Supplementary Video 1), and from astrocytes to metastatic cells by flow cytometry (Extended Data Fig. 1d). Astrocyte calcein transfer occurred more readily with brain metastatic cells than with their parental counterparts (Fig. 1e). This phenotype was not fully explained by brain metastatic cell expression of Cx43 (Fig. 1f, Extended Data Fig. 2a, b) or other astrocytic connexins (Cx26, Cx30) (Extended Data Fig. 2c). Cx43 expression was higher in astrocytes than in brain metastatic cells (Fig. 1g, Extended Data Fig. 2d).

Reasoning that cancer cells must use another component besides Cx43 to engage astrocytes, we investigated the cadherin-related

¹Cancer Biology and Genetics Program, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. ²Department of Neurology, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. ³Donald B. and Catherine C. Marron Cancer Metabolism Center, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. ⁴Molecular Cytology Core Facility, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. [†]Present addresses: The Wistar Institute 3601 Spruce Street, Philadelphia, Pennsylvania 19104, USA (Q.C.); Cancer Program, The Eli and Edythe L. Broad Institute, Cambridge, Massachusetts 02142, USA (X.J.); Brain Metastasis Group, Spanish National Cancer Research Centre (CNIO), Madrid E28029, Spain (M.V.); Department of Functional Biology IUOPA, University of Oviedo, Facultad de Medicina, 33006 Oviedo, Spain (A.L.-S.); Department of Genetics, Beth Israel Deaconess Medical Center, Harvard Medical School, 3 Blackfan Circle, CLS 417, Boston, Massachusetts 02115, USA (L.J.).

*These authors contributed equally to this work.

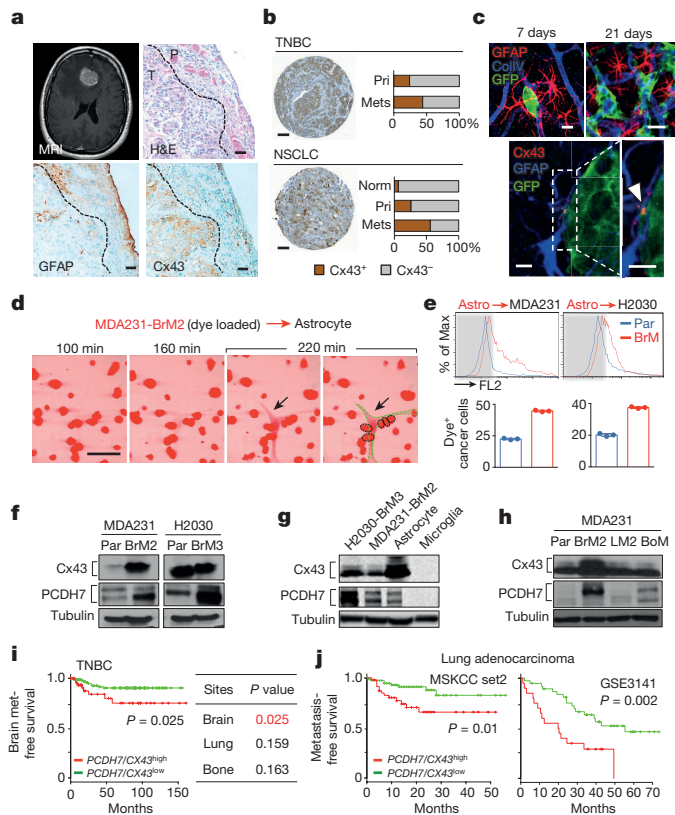


Figure 1 | Cx43 and PCDH7 are associated with brain metastasis.

a, Top left, contrast-enhanced MRI of representative patient with brain metastasis. Tumour (white) is surrounded by parenchymal reaction (dark grey). Top right, haematoxylin and eosin (H&E) staining of resected brain metastasis (T) and parenchyma (P). Bottom, immunohistochemistry of adjacent sections for GFAP (left) and Cx43 (right). Scale bars, 10 μ m; $n = 6$ patient samples. **b**, Cx43 expression is increased in brain metastases (Mets) compared with primary (Pri) and normal (Norm) tissue. Representative images of Cx43 staining in clinical samples from triple-negative breast cancer (TNBC) and non-small cell lung carcinoma (NSCLC). Proportion of Cx43-positive samples was quantified in primary tumours (TNBC $n = 98$, NSCLC $n = 138$), brain metastases (TNBC $n = 117$; NSCLC $n = 91$) and normal lung tissues ($n = 75$). Scale bars, 100 μ m. **c**, Top, GFP⁺ H2030-BrM3 cells (green) are surrounded by GFAP⁺ activated astrocytes (red) in the brain parenchyma at early (day 7) and later (day 21) time points after intracardiac inoculation in mice. Blue, collagen IV (ColIV) staining in vessels. Scale bars, 10 μ m. Bottom, Cx43 staining (arrowhead) at the interface of GFP⁺ H2030-BrM3 (green) and GFAP⁺ astrocytes (blue). Scale bars, 10 μ m. **d**, Time-lapse images of dye transfer from MDA231-BrM2 cells to astrocytes. See also Supplementary Video 1. Scale bar, 100 μ m. **e**, Quantification of dye transfer from astrocytes to cancer cells. Histograms show red fluorescent signal in parental (Par) and BrM cells. FL2 denotes emitted fluorescent light of calcein. Data are mean \pm s.e.m. (from $n = 3$ biological replicates over 3 independent experiments). **f–h**, Cx43 and PCDH7 western blotting in the indicated parental and brain metastatic derivatives (**f**, $n = 3$ independent experiments), in brain metastatic cells compared to brain cell types (**g**, $n = 2$ independent experiments), and in the MDA231 derivatives metastatic to brain, lung (LM) or bone (BoM) (**h**, $n = 2$ independent experiments). Full blots are shown in Supplementary Data. **i, j**, Kaplan–Meier plots of brain metastasis-free survival in 189 cases of TNBC (**i**) and 129 cases (MSKCC set2) and 58 cases (GSE3141) of lung adenocarcinoma (**j**), on the basis of Cx43 and PCDH7 expression in the primary tumour.

neuronal receptor PCDH7, encoded by *PDCH7*, one of a small group of genes upregulated in brain metastatic cells from both breast and lung tumours^{3,6,12}. Protocadherins are integral membrane proteins that direct cell–cell contacts by homophilic interaction, and PCDH7 is the

only protocadherin expressed predominantly in the brain^{15,16}. PCDH7 levels were higher in brain metastatic derivatives than in parental cell lines (Fig. 1f, Extended Data Fig. 2a, b), and higher than in matched derivatives highly metastatic to bone or lung but not brain (Fig. 1h; refer to Extended Data Fig. 1b). The PCDH7 level in brain metastatic cells was higher than in astrocytes, microglia or endothelial cells (Fig. 1g, Extended Data Fig. 2d).

In clinical cohorts of TNBC, the expression of *PCDH7* and *CX43* (also known as *GJA1*) in primary tumours was associated with brain, but not bone or lung metastasis (Fig. 1i). Up to 70% of relapses in patients with NSCLC include brain metastases¹⁷; contributing disproportionately to survival¹⁸. *CX43* and *PCDH7* expression was associated with decreased metastasis-free survival of NSCLC patients in three cohorts (Fig. 1j, Extended Data Fig. 2e). These results all support the relevance of PCDH7 and Cx43 in brain metastasis.

PCDH7 directs carcinoma–astrocyte gap junctions

Brain metastatic cells depleted of either PCDH7 or Cx43 using short hairpin RNAs (shRNAs) (Extended Data Fig. 2f, g) showed reduced capacity for dye transfer to astrocytes (Fig. 2a, Extended Data Fig. 3a), on par with the pan-connexin inhibitor, carbenoxolone (Extended Data Fig. 3b). Cadherins may establish homophilic binding between adjacent cells¹⁹. We proposed that astrocyte PCDH7 might participate in the formation of gap junctions with brain metastatic cancer cells, similar to gap junctions between astrocytomas and neighbouring astrocytes^{20,21}. Indeed, PCDH7 depletion in astrocytes (Extended Data Fig. 3c) inhibited dye transfer from MDA231-BrM2 cells (Extended Data Fig. 3d).

Human brain microvascular endothelial cells (HBMECs) have no detectable PCDH7 expression and express low levels of Cx43 (Extended Data Fig. 2d). Despite low gap junction communication between cancer cells and HBMECs (Extended Data Fig. 3e), dye transfer between cancer cell and astrocyte was favoured over dye transfer between cancer cell and HBMECs (Extended Data Fig. 3f). Primary microglia cells expressed very low levels of Cx43 and PCDH7 (Fig. 1g) and did not accept calcein from cancer cells (Extended Data Fig. 3g). Thus, PCDH7 directs cancer cells to form Cx43 gap junctions with astrocytes preferentially.

To detect Cx43–PCDH7 interactions in live cells, we used a split luciferase complementation assay²². Constructs encoding *PDCH7* and *CX43* fused to the N-terminal (NLuc) and C-terminal (CLuc) halves of firefly luciferase were expressed in non-green fluorescent protein (GFP)-luciferase-labelled parental cells. When NLuc and CLuc cytoplasmic domains come into proximity, luciferase activity is reconstituted (Extended Data Fig. 4a). Cx43 self-assembly served as positive control. We detected specific luciferase activity in cells expressing both Cx43-CLuc and PCDH7-NLuc (Extended Data Fig. 4b), consistent with Cx43 and PCDH7 interaction within the same cell. The expression levels of PCDH7 and Cx43 luciferase chimaeras were comparable to endogenous levels (Extended Data Fig. 4c). Astrocytes increased luciferase signal in cancer cells after co-culture (Extended Data Fig. 4d), suggesting that astrocyte Cx43 and PCDH7 induce further clustering of cancer cell Cx43-CLuc and PCDH7-NLuc. No interaction was detected between Cx43-CLuc and N-cadherin or E-cadherin fused with NLuc (Extended Data Fig. 4e–g).

Cx43 and PCDH7 mediate brain metastasis

shRNA-mediated depletion of either Cx43 or PCDH7 in breast cancer and lung cancer cells inhibited brain metastases growth in both immunocompetent (Fig. 2b) and xenograft models, (Fig. 2c, d and Extended Data Fig. 5a, b); this did not affect the formation of lung lesions (Extended Data Fig. 5c).

The Cx43(T154A) mutant assembles hemichannels but lacks channel function²³. Cx43(T154A) re-expressed in Cx43-depleted brain metastatic cancer cells (Extended Data Fig. 5d) was unable to mediate calcein transfer from astrocyte to MDA231-BrM cells (Fig. 2e).

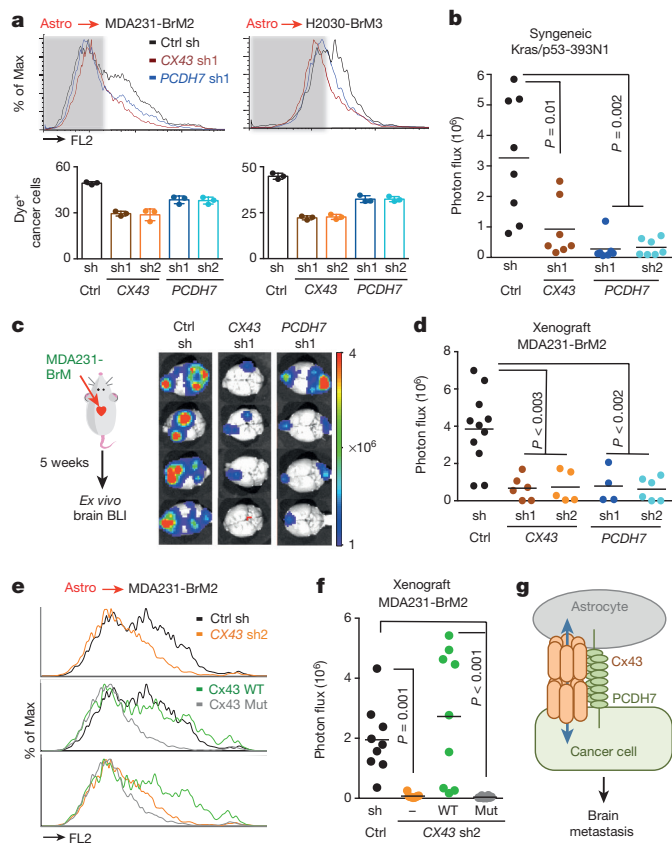


Figure 2 | Cx43-PCDH7 carcinoma-astrocyte gap junctions mediate brain metastasis. **a**, Histograms (top) and quantification (bottom) of dye transfer from astrocytes to control and Cx43-depleted or PCDH7-depleted brain metastatic cells. Data are mean \pm s.e.m. (from $n = 3$ biological replicates over 3 independent experiments). **b–d**, Bio-luminescent imaging (BLI) (**c**) and quantification (**b**, **d**) of brain metastatic lesions formed by control (Ctrl), Cx43-depleted (CX43 sh1), or PCDH7-depleted (PCDH7 sh1/2) brain metastatic cells in the MDA231 xenograft model or 393N1 syngeneic models of brain metastasis ($n = 3$ independent experiments, $n = 8–10$ mice per group). All source data from mouse experiments are in Supplementary Information. **e**, **f**, Wild type (WT) or T154A mutant (Mut) Cx43 was re-expressed in Cx43-depleted (CX43 sh2) MDA231-BrM2 cells. Cells were subjected to astrocyte dye transfer analysis by flow cytometry (**e**, $n = 3$ independent experiments), or to brain metastasis assays and BLI quantification (**f**, $n = 2$ independent experiments, 9 mice per group). **g**, Schematic summary of Cx43- and PCDH7-mediated interactions between cancer cells and astrocytes in brain metastasis.

Wild-type Cx43 rescued brain metastatic activity in Cx43-depleted MDA231-BrM and H2030-BrM cells, whereas Cx43(T154A) did not (Fig. 2f, Extended Data Fig. 5e). These observations support a model in which PCDH7 directly interacts with Cx43 to assemble functional gap junctions between cancer cells and astrocytes (Fig. 2f).

We defined the stage at which PCDH7 and Cx43 contribute to the formation of brain metastases using short-term metastasis assays with MDA231-BrM2 cells. In this model, extravasation across the BBB is complete 7 days after inoculation, vascular co-option and overt outgrowth occur by day 14 (ref. 6). Cx43 or PCDH7 depletion in the cancer cells did not significantly diminish the number of GFP⁺ cancer cells in the brain parenchyma at day 7 (Extended Data Fig. 6a). Fourteen days after inoculation, micrometastases resulting from Cx43- or PCDH7-depleted cells showed decreased proliferation (Extended Data Fig. 6b). In *ex vivo* brain slice assays⁶, Cx43- or PCDH7-depleted cells displayed increased cleaved caspase 3 staining (Extended Data Fig. 6c) and maintained vascular contacts (Extended Data Fig. 6d). Thus, cancer cell-astrocyte gap junction channels support brain metastasis growth after extravasation and vascular cooption.

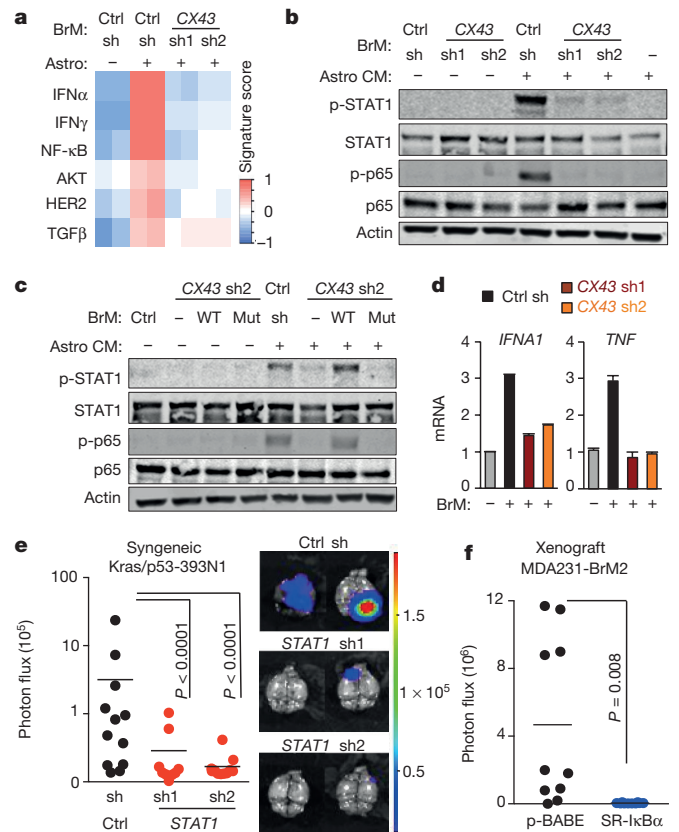


Figure 3 | Gap junctions activate STAT1 and NF- κ B pathways in cancer cells. **a**, Signalling pathway analysis of TRAP RNA-seq data from MDA231-BrM2 cells after co-culture with astrocytes. Control or Cx43-depleted MDA231-BrM2 cells expressing an eGFP-L10a ribosomal protein fusion were co-cultured with astrocytes for 24 h before polysome immunoprecipitation and mRNA sequencing. Heatmap depicts blue (downregulated) and red (upregulated) pathways. $n = 2$ biological replicates. **b**, **c**, STAT1 and NF- κ B p65 phosphorylation (p-p65) in MDA231-BrM2 cells after a 2 h incubation with conditioned media (CM) from indicated cancer cell/astrocyte co-culture. Media samples were collected after 24 h co-culture of astrocytes with control or Cx43-depleted MDA231-BrM2 cells (**b**), or from Cx43-depleted MDA231-BrM2 cells that were transduced with wild-type Cx43 or Cx43(T154A) mutant (**c**) ($n \geq 3$ independent experiments in **b** and **c**). **d**, Relative mRNA levels of *IFNA1* and *TNF* in human astrocytes re-isolated after co-culture with MDA231-BrM2 cancer cells. Data are mean \pm s.e.m. (from $n = 3$ biological replicates in 2 independent experiments). **e**, **f**, Quantification of BLI signal from brain metastases formed by control or STAT1-knockdown (STAT1 sh1/2) Kras/p53-393N1 cells (**e**), and SR-1xB α MDA231-BrM2 cells (**f**) ($n = 2$ independent experiments, with 12 mice total per group in **e** and 10 mice total per group in **f**).

Cancer cells trigger astrocyte cytokine release

To determine the mechanism behind Cx43-mediated brain metastatic growth, we assayed cancer cell gene expression by translating ribosome affinity purification (TRAP)²⁴ (Extended Data Fig. 7a). We expressed the enhanced GFP (eGFP)-tagged L10a ribosomal subunit in MDA231-BrM2 cells with either basal or reduced Cx43 expression. After cancer cell-astrocyte co-culture, we immunoprecipitated eGFP-tagged polysomes and sequenced the associated mRNA (Extended Data Fig. 7b, c). The gene expression patterns revealed that IFN and NF- κ B pathways in brain metastatic cells were activated after astrocyte co-culture in a Cx43-dependent manner (Fig. 3a). Conditioned media from the co-cultures was sufficient to activate the IFN and NF- κ B signalling in the cancer cells, as determined by phosphorylation of STAT1 and NF- κ B p65 (Fig. 3b, Extended Data Fig. 7d). Cx43 channel function was required for this effect (Fig. 3c).

IFN α , TNF and TGF α accumulated in conditioned media from MDA231-BrM2 cell-astrocyte co-cultures in a gap-junction-dependent manner (Extended Data Fig. 8a, b). MDA231-BrM2 cells, either alone or co-cultured with astrocytes, did not express these cytokines by TRAP RNA-sequencing (RNA-seq; data not shown). Upregulation of *TNF* and *IFNA1* mRNA was detected in astrocytes re-isolated after co-culture (Fig. 3d). Addition of IFN α or TNF inhibited brain metastatic cancer cell apoptosis in response to chemotherapy (Extended Data Fig. 8c). In two syngeneic mouse models, knockdown of STAT1 in brain metastatic cells (Extended Data Fig. 8d) reduced brain metastasis (Fig. 3e, Extended Data Fig. 8e). Inhibition of NF- κ B by overexpression of I κ B α super suppressor (SR-I κ B α)²⁵ in brain metastatic cells (Extended Data Fig. 8f) also suppressed brain metastasis (Fig. 3f). These results suggest that heterocellular gap junction communication elicits production of IFN α and TNF in astrocytes, triggering STAT1 and NF- κ B survival signals in cancer cells.

Cancer cells transfer cGAMP to astrocytes

Upregulation of both IFN α and TNF was reminiscent of a cellular response to dsDNA²⁶. Cytosolic dsDNA activates the cGAS-STING pathway, an innate immune response against viral infection²⁷, in which cyclic GMP-AMP synthase (cGAS) senses cytosolic dsDNA and synthesizes the second messenger 2'3'-cyclic GMP-AMP (cGAMP). cGAMP binding to STING triggers phosphorylation and activation of TBK1 and IRF3, nuclear accumulation of IRF3, and transcriptional activation of IRF3 target genes *IFNA1* and *TNF*¹⁰.

Co-incubation of MDA231-BrM2 cells and astrocytes triggered Cx43-dependent phosphorylation of TBK1 and IRF3 (Fig. 4a, Extended Data Fig. 9a). Nuclear accumulation of IRF3 occurred only in co-cultured astrocytes, and not in either cell type alone (Fig. 4b). STING knockdown in mouse astrocytes (Extended Data Fig. 9b) inhibited their ability to respond to mouse LLC-BrM cells with IRF3 phosphorylation (Fig. 4c), or IFN α and TNF production (Extended Data Fig. 9c), indicating the need for STING activity in astrocytes. We inoculated LLC-BrM cells into syngeneic STING-mutant or wild-type C57Bl6 mice and found that host STING inactivation suppressed brain metastasis by these cells (Fig. 4d, Extended Data Fig. 9d).

We next investigated the source of astrocyte STING activation. Subcellular fractionation (Extended Data Fig. 9e, f) and immunofluorescence (Extended Data Fig. 9g) showed cytosolic dsDNA in human cancer cell lines and not in astrocytes and other non-neoplastic cells. We detected cGAMP by liquid chromatography tandem mass spectrometry (LC-MS/MS) in MDA231-BrM2 cells, but not in astrocytes (Fig. 4e, Extended Data Fig. 9h). Co-culture of MDA231-BrM2 cells with astrocytes further increased cGAMP levels in a Cx43-dependent manner (Fig. 4f). Given the presence of both cytosolic dsDNA and cGAMP in cancer cells (Extended Data Fig. 9i, j), we next identified which molecule was responsible for astrocyte STING pathway activation. cGAS knockdown cancer cells (Extended Data Fig. 9k) induced little IRF3 phosphorylation (Fig. 4g), TNF or IFN α production (Extended Data Fig. 9l) in co-cultured astrocytes. Moreover, cGAS-depletion in cancer cells led to reduced brain metastasis (Extended Data Fig. 9m).

Together, these results support a model in which brain metastatic cancer cells contain cytosolic dsDNA and cGAMP and engage astrocytes in Cx43-based gap junctions. The gap junctions allow passage of cGAMP from cancer cells into astrocytes to trigger TBK1 and IRF3 activation and production of IFN α and TNF. These cytokines activate STAT1 and NF- κ B signalling in cancer cells to support cancer cell growth and survival under microenvironmental and chemotherapeutic stresses (Fig. 4h).

Gap junction directed therapy

Genetic evidence that inhibition of gap junction signalling decreases brain metastasis prompted testing pharmacological suppressors of gap junction activity. In addition to anti-inflammatory activity, meclofenamate inhibits Cx43 gap junction gating²⁸ and inhibits

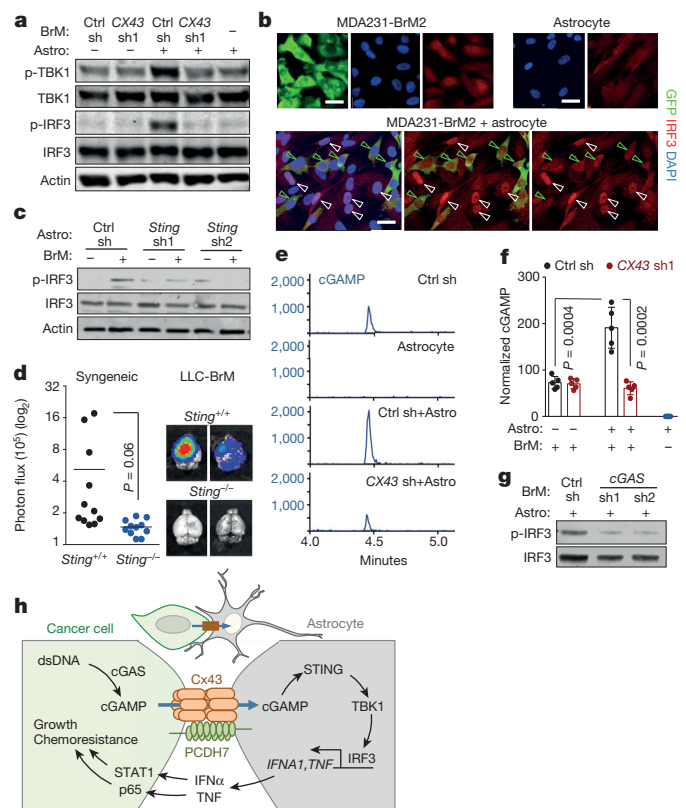


Figure 4 | Gap junctions induce cytosolic dsDNA response in astrocytes.

a, MDA231-BrM2 cells expressing control or CX43 shRNA were cultured for 18 h with/without astrocytes, and subjected to immunoblotting analysis of phosphorylated (p-) TBK1 and IRF3 ($n = 3$ independent experiments). **b**, Representative images of dual immunofluorescent staining of IRF3 and GFP. DAPI denotes nuclear staining. White arrowheads denote nuclear accumulation of IRF3 in astrocytes; green arrowheads denote even distribution of IRF3 in GFP⁺ MDA231-BrM2 cells. Scale bars, 20 μ m; $n = 2$ independent experiments. **c**, Mouse astrocytes (Astro) expressing control (non-silencing) shRNA or shRNA targeting *Sting* (also known as *Tmem173*), were cultured for 18 h with/without LLC-BrM cells, and subjected to immunoblotting analysis of p-IRF3. **d**, LLC-BrM growth in syngeneic C57Bl6 mice hosts wild-type (+/+) or knockout (-/-) for *Sting*. Right, quantification of BLI signal from brain metastases formed in *Sting*^{+/+} and *Sting*^{-/-} host mice ($n = 11$ mice in each group; 2 independent experiments). **e**, **f**, MDA231-BrM2 alone, astrocytes alone, or 18 h co-cultures, were collected for sample preparation and cGAMP analysis by LC-MS/MS. Representative chromatograms (**e**) and quantification (**f**) are shown ($n = 5$ biological replicates in 3 independent experiments). See also Supplementary Information. **g**, Human astrocytes, were cultured for 18 h with or without H2030-BrM cells (BrM) expressing control shRNA or shRNA targeting *cGAS* (also known as *MB21D1*), and subjected to immunoblotting analysis of p-IRF3 (2 independent experiments). **h**, Schematic summary of gap-junction-mediated anti-dsDNA response, production of IFN α and TNF in astrocytes, and consequent activation of STAT1 and NF- κ B pathways in cancer cells to support brain metastasis.

epileptogenesis in animal models²⁹. Tonabersat is a benzopyran derivative that binds to a unique stereoselective binding site in astrocytes^{30,31} and inhibits gap-junction-mediated processes including cortical spreading depression³² and trigeminal ganglion neuronal-satellite cell signalling³³. Both tonabersat and meclofenamate inhibited dye transfer from astrocytes to cancer cells (Fig. 5a) and release of IFN α and TNF in astrocyte cancer cell co-cultures (Fig. 5b). Treatment with either compound inhibited brain metastases in xenograft and immunocompetent models (Fig. 5c, Extended Data Fig. 10a, b), but did not restrict lung metastasis (Extended Data Fig. 10c, d). Neither drug altered the astrocyte response to cGAMP (Extended Data Fig. 10e).

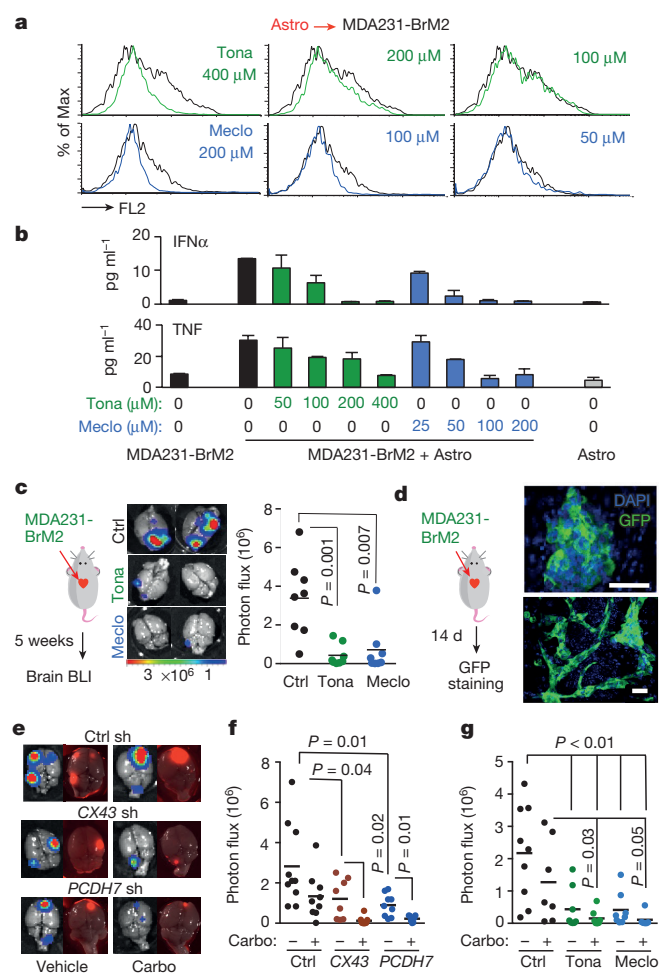


Figure 5 | Inhibition of gap junction activity controls brain metastatic outgrowth. **a**, Dye transfer from astrocytes to MDA231-BrM2 cells in the presence of the indicated concentrations of tonabersat (Tona) or meclofenamate (Meclo) ($n \geq 3$ independent experiments). **b**, ELISA of IFN α and TNF in conditioned media from co-cultured MDA231-BrM2 cell and astrocytes in the presence of tonabersat (50–400 μ M) or meclofenamate (25–200 μ M). All graphs show mean \pm s.e.m. (from 2 independent experiments with 4 replicates each). **c**, Tonabersat or meclofenamate was administered daily starting one day after cancer cell inoculation in mice. Brain metastatic lesions were quantified based on BLI ($n = 2$ independent experiments; 8 mice total per group). **d**, GFP staining of 14-day brain metastatic lesions. Representative images show large, progressive lesions. DAPI, nuclear staining. Scale bars, 40 μ m; $n = 10$ mice. **e–g**, Fourteen days after inoculation with MDA231-BrM2 cells transduced with inducible control, CX43 or PCDH7 shRNAs, mice were treated with doxycycline and carboplatin. **e**, Representative images of matched *ex vivo* brain BLI and red fluorescence imaging. **f**, Brain metastatic lesions were quantified based on BLI (2 independent experiments with $n = 10$ total mice per group). **g**, Fourteen days after inoculation with MDA231-BrM2 cells, mice were treated with tonabersat, meclofenamate, and carboplatin. Following the indicated regimens, brain metastatic lesions were quantified based on BLI (2 independent experiments with $n = 9$ mice total per group).

To test the effect of CX43 or PCDH7 depletion in established metastases, we transduced MDA231-BrM2 cells with doxycycline-inducible shRNA expression vectors (Extended Data Fig. 10g). Red fluorescence protein (RFP) under the control of the same promoter provided a marker of hairpin expression (Extended Data Fig. 10f). Doxycycline treatment began 14 days after inoculation^{3,6} (Extended Data Fig. 10h, Fig. 5d), and reduced brain metastatic burden 3 weeks later (Fig. 5e, f).

Brain metastases show pronounced resistance to chemotherapy³⁴. Carboplatin crosses the BBB³⁵, modestly improving overall survival in patients with brain metastases from breast³⁶ or lung cancer³⁷. Carboplatin alone (50 mg kg⁻¹ every fifth day) starting on day 14 inhibited brain metastasis to a similar extent to CX43 or PCDH7 depletion (Fig. 5e, f); the combination of carboplatin and CX43- or PCDH7-depletion further reduced the metastatic burden (Fig. 5e, f). Treatment with either tonabersat (10 mg kg⁻¹) or meclofenamate (20 mg kg⁻¹) as single agents (Fig. 5g) significantly inhibited progression of metastatic lesions. Addition of carboplatin to either agent profoundly inhibited brain metastasis (Fig. 5g).

Discussion

The brain represents a formidable metastatic target, of which astrocytes are a predominant feature. We find that cancer cells use PCDH7 to engage astrocytes selectively in vital CX43 gap junctions. Cadherins mediate cell–cell communication in development and tissue homeostasis¹⁹, particularly in the nervous system³⁸. Remarkably, brain metastatic cells adopt a cadherin whose normal expression is largely confined to the brain¹⁵. PCDH7 joins the sialyltransferase ST6GALNAC5 (ref. 3) and neuroserpin⁶ as brain-restricted components that metastatic cells from breast and lung carcinomas selectively express to colonize the brain. Functional CX43-based gap junctions between cancer cells and astrocytes allow cancer cells to disseminate cGAMP to astrocytes. This activates the astrocytic cGAS-STING pathway and release of cytokines including IFN α and TNF, which provide a growth advantage for brain metastatic cells by protecting against physiological and chemotherapeutic stresses.

cGAMP transfer from cancer cell to astrocyte is reminiscent of cGAMP spread to adjacent cells in the anti-viral context^{39,40}. However, unlike homotypic transfer of cGAMP to bystander cells to intensify the immune response, brain metastatic cells shunt cGAMP into neighbouring host astrocytes to trigger downstream signalling that supports metastatic outgrowth. This pro-metastatic process contrasts with previous reports of tumour STING activation and subsequent host immune cell extracranial anti-tumour response^{41,42} and highlights the profound impact of stromal context.

Brain metastases are a major contributor to cancer mortality, with few therapeutic options available. Cancer cell dependency on CX43–PCDH7 gap junctions for survival, and outgrowth of metastatic lesions suggests a therapeutic opportunity. Our pre-clinical results using combinations of chemotherapy and gap junction modulators provide a proof-of-principle for the therapeutic potential of these interventions against brain metastasis.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 September 2015; accepted 12 April 2016.

Published online 18 May 2016.

- Gavrilovic, I. T. & Posner, J. B. Brain metastases: epidemiology and pathophysiology. *J. Neurooncol.* **75**, 5–14 (2005).
- Stelzer, K. J. Epidemiology and prognosis of brain metastases. *Surg. Neurol. Int.* **4**, S192–S202 (2013).
- Bos, P. D. *et al.* Genes that mediate breast cancer metastasis to the brain. *Nature* **459**, 1005–1009 (2009).
- Eichler, A. F. *et al.* The biology of brain metastases—translation to new therapies. *Nature Rev. Clin. Oncol.* **8**, 344–356 (2011).
- Kienast, Y. *et al.* Real-time imaging reveals the single steps of brain metastasis formation. *Nature Med.* **16**, 116–122 (2010).
- Valiente, M. *et al.* Serpins promote cancer cell survival and vascular co-option in brain metastasis. *Cell* **156**, 1002–1016 (2014).
- Giaume, C., Koulakoff, A., Roux, L., Holcman, D. & Rouach, N. Astroglial networks: a step further in neuroglial and gliovascular interactions. *Nature Rev. Neurosci.* **11**, 87–99 (2010).
- Sofroniew, M. V. & Vinters, H. V. Astrocytes: biology and pathology. *Acta Neuropathol.* **119**, 7–35 (2010).
- Kim, S. J. *et al.* Astrocytes upregulate survival genes in tumor cells and induce protection from chemotherapy. *Neoplasia* **13**, 286–298 (2011).

10. Wu, J. *et al.* Cyclic GMP-AMP is an endogenous second messenger in innate immune signaling by cytosolic DNA. *Science* **339**, 826–830 (2013).
11. Theis, M. & Giaume, C. Connexin-based intercellular communication and astrocyte heterogeneity. *Brain Res.* **1487**, 88–98 (2012).
12. Nguyen, D. X. *et al.* WNT/TCF signaling through LEF1 and HOXB9 mediates lung adenocarcinoma metastasis. *Cell* **138**, 51–62 (2009).
13. Winslow, M. M. *et al.* Suppression of lung adenocarcinoma progression by Nkx2-1. *Nature* **473**, 101–104 (2011).
14. Oshima, A. Structure and closure of connexin gap junction channels. *FEBS Lett.* **588**, 1230–1237 (2014).
15. Yoshida, K., Yoshitomo-Nakagawa, K., Seki, N., Sasaki, M. & Sugano, S. Cloning, expression analysis, and chromosomal localization of BH-protocadherin (PCDH7), a novel member of the cadherin superfamily. *Genomics* **49**, 458–461 (1998).
16. Kim, S. Y., Chung, H. S., Sun, W. & Kim, H. Spatiotemporal expression pattern of non-clustered protocadherin family members in the developing rat brain. *Neuroscience* **147**, 996–1021 (2007).
17. Gaspar, L. E. *et al.* Time from treatment to subsequent diagnosis of brain metastases in stage III non-small-cell lung cancer: a retrospective review by the Southwest Oncology Group. *J. Clin. Oncol.* **23**, 2955–2961 (2005).
18. Gaspar, L. E., Scott, C., Murray, K. & Curran, W. Validation of the RTOG recursive partitioning analysis (RPA) classification for brain metastases. *Int. J. Radiat. Oncol. Biol. Phys.* **47**, 1001–1006 (2000).
19. Yagi, T. & Takeichi, M. Cadherin superfamily genes: functions, genomic organization, and neurologic diversity. *Genes Dev.* **14**, 1169–1180 (2000).
20. Osswald, M. *et al.* Brain tumour cells interconnect to a functional and resistant network. *Nature* **528**, 93–98 (2015).
21. Sin, W. C. *et al.* Astrocytes promote glioma invasion via the gap junction protein connexin43. *Oncogene* **35**, 1504–1516 (2015).
22. Luker, K. E. *et al.* Kinetics of regulated protein-protein interactions revealed with firefly luciferase complementation imaging in cells and living animals. *Proc. Natl Acad. Sci. USA* **101**, 12288–12293 (2004).
23. Beahm, D. L. *et al.* Mutation of a conserved threonine in the third transmembrane helix of α - and β -connexins creates a dominant-negative closed gap junction channel. *J. Biol. Chem.* **281**, 7994–8009 (2006).
24. Heiman, M. *et al.* A translational profiling approach for the molecular characterization of CNS cell types. *Cell* **135**, 738–748 (2008).
25. Boehm, J. S. *et al.* Integrative genomic approaches identify IKBKE as a breast cancer oncogene. *Cell* **129**, 1065–1079 (2007).
26. Cai, X., Chiu, Y. H. & Chen, Z. J. The cGAS-cGAMP-STING pathway of cytosolic DNA sensing and signaling. *Mol. Cell* **54**, 289–296 (2014).
27. Stetson, D. B. & Medzhitov, R. Recognition of cytosolic DNA activates an IRF3-dependent innate immune response. *Immunity* **24**, 93–103 (2006).
28. Harks, E. G. *et al.* Fenamates: a novel class of reversible gap junction blockers. *J. Pharmacol. Exp. Ther.* **298**, 1033–1041 (2001).
29. Jin, M. *et al.* Effects of meclofenamic acid on limbic epileptogenesis in mice kindling models. *Neurosci. Lett.* **543**, 110–114 (2013).
30. Chan, W. N. *et al.* Identification of (–)-cis-6-acetyl-4S-(3-chloro-4-fluoro-benzoylamino)-3,4-dihydro-2,2-dimethyl-2H-benzo[b]pyran-3S-ol as a potential antimigraine agent. *Bioorg. Med. Chem. Lett.* **9**, 285–290 (1999).
31. Herdon, H. J. *et al.* Characterization of the binding of [³H]-SB-204269, a radiolabelled form of the new anticonvulsant SB-204269, to a novel binding site in rat brain membranes. *Br. J. Pharmacol.* **121**, 1687–1691 (1997).
32. Read, S. J., Smith, M. I., Hunter, A. J., Upton, N. & Parsons, A. A. SB-220453, a potential novel antimigraine agent, inhibits nitric oxide release following induction of cortical spreading depression in the anaesthetized cat. *Cephalalgia* **20**, 92–99 (2000).
33. Damodaram, S., Thalakoti, S., Freeman, S. E., Garrett, F. G. & Durham, P. L. Tonabersat inhibits trigeminal ganglion neuronal-satellite glial cell signaling. *Headache* **49**, 5–20 (2009).
34. Deeken, J. F. & Loscher, W. The blood-brain barrier and cancer: transporters, treatment, and Trojan horses. *Clin. Cancer Res.* **13**, 1663–1674 (2007).
35. Pitz, M. W., Desai, A., Grossman, S. A. & Blakeley, J. O. Tissue concentration of systemically administered antineoplastic agents in human brain tumors. *J. Neurooncol.* **104**, 629–638 (2011).
36. Lim, E. & Lin, N. U. Updates on the management of breast cancer brain metastases. *Oncology* **28**, 572–578 (2014).
37. Taimur, S. & Edelman, M. J. Treatment options for brain metastases in patients with non-small-cell lung cancer. *Curr. Oncol. Rep.* **5**, 342–346 (2003).
38. Hirano, S., Suzuki, S. T. & Redies, C. The cadherin superfamily in neural development: diversity, function and interaction with other molecules. *Front. Biosci.* **8**, d306–d355 (2003).
39. Patel, S. J., King, K. R., Casali, M. & Yarmush, M. L. DNA-triggered innate immune responses are propagated by gap junction communication. *Proc. Natl Acad. Sci. USA* **106**, 12867–12872 (2009).
40. Ablasser, A. *et al.* Cell intrinsic immunity spreads to bystander cells via the intercellular transfer of cGAMP. *Nature* **503**, 530–534 (2013).
41. Demaria, O. *et al.* STING activation of tumor endothelial cells initiates spontaneous and therapeutic antitumor immunity. *Proc. Natl Acad. Sci. USA* **112**, 15408–15413 (2015).
42. Woo, S. R. *et al.* STING-dependent cytosolic DNA sensing mediates innate immune recognition of immunogenic tumors. *Immunity* **41**, 830–842 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank D. Macalino and other members of the Massagué laboratory for discussions. This work was supported by NIH grants P01-CA129243, U54-163167 and P30 CA008748, DOD Innovator award W81XWH-12-0074, the Alan and Sandra Gerry Metastasis Research Initiative (J.M.), the MSKCC Clinical Scholars Training Program (A.B.), the Solomon R. and Rebecca D. Baker Foundation (A.B.), and by the Susan G. Komen Organization (X.J.).

Author Contributions Q.C., A.B. and J.M. conceptualized the project and designed the experiments. Q.C. and A.B. performed the experiments. X.J., M.V., E.E.E., A.L.-S., L.J. and R.P. assisted with the experiments and bioinformatics analysis. H.S. and J.R.C. performed the LC-MS/MS analysis, and K.X. the time-lapse confocal imaging. A.B., Q.C. and J.M. wrote the paper.

Author Information RNA-seq data have been deposited in NCBI Gene Expression Omnibus (GEO) under accession number GSE79256. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.M. (j-massague@ski.mskcc.org).

Reviewer Information Nature thanks R. Hynes and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Cell culture. Human MDA-MB-231, mouse MMTV-neu, and mouse Lewis lung carcinoma (LLC) and their metastatic derivatives, MDA-MB-231-BrM2 (MDA231-BrM), ErbB-BrM, and LLC-BrM respectively, and the mouse 373N1, 393N1, 482N1 and 2691N1 cell lines were cultured in DMEM with 10% FBS and 2 mM L-glutamine. Human H2030 cells and metastatic derivatives H2030-BrM3 (H2030-BrM) were cultured in RPMI 1640 medium supplemented with 10% FBS and 2 mM L-glutamine. For lentivirus production, 293T cells were cultured in DMEM supplemented with 10% FBS and 2 mM L-glutamine. Human primary astrocytes, mouse primary astrocytes, HBMECs, adult dermal fibroblasts, and microglia were cultured in media specified by the supplier (ScienCell), and used between passages 2 and 6. All cells tested negative for mycoplasma.

Animal studies. All experiments using animals were performed in accordance with protocols approved by the MSKCC Institutional Animal Care and Use Committee. Athymic NCR nu/nu mice (NCI-Frederick), Cr:NIH bg-nu-xid mice (NCI-Frederick), B6129SF1/J, C57BL/6J-*Tmem173gt^{f43}* 'golden ticket', and C57/BL/6J mice (Jackson Laboratory) were used at 5–6 weeks of age. For long-term brain metastasis assays we followed previously described procedures³. In brief, 1×10^4 MDA231-BrM2 cells, 5×10^4 H2030-BrM3 cells, 5×10^4 LLC-BrM or 1×10^5 393N1 cells suspended in 100 μ l PBS were injected into the left cardiac ventricle. At the experimental endpoint, anaesthetized mice (ketamine 100 mg kg⁻¹, xylazine 10 mg kg⁻¹) were injected retro-orbitally with D-luciferin (150 mg kg⁻¹), before euthanasia, dissection of brain and quantified by *ex vivo* BLI. For short-term (7-day and 14-day) brain metastasis experiments, we introduced 5×10^5 cells. Where relevant, TRITC dextran (70 kDa) (Life Technologies) was intravenously injected to stain vascular structures before dissection. For lung colonization assays, 2×10^5 MDA231-BrM2 cells in 100 μ l PBS were injected into the lateral tail vein. For orthotopic tumour implantation, 5×10^3 cells in 50 μ l of 1:1 mix of PBS/growth-factor-reduced Matrigel (BD Biosciences) were injected into the fourth right mammary fat pad of female mice. For inducible knockdown experiments, mice were given doxycycline hyclate (Sigma-Aldrich) in the drinking water (2 mg ml⁻¹) and the diet (Harlan) 14 days after injection of cancer cells. For drug treatment experiments, mice were intraperitoneally injected with carboplatin (Hospira) (5 mg kg⁻¹ per 5 days), tonabersat (MedChem Express) (10 mg kg⁻¹ day⁻¹), or meclofenamic acid sodium salt (Sigma-Aldrich) (20 mg kg⁻¹ day⁻¹). Vehicle (10% DMSO in polyethylene glycol 400) was used in control mice. Quantification of tumour burden was by BLI, performed using an IVIS Spectrum Xenogen instrument (Caliper Life Sciences) and analysed using Living Image software v.2.50. See Source Data. A priori sample size determination for animal experiments was determined by Mead's Resource Equation: 10 animals per treatment group in an experimental design of three groups, without further stratification, gives 28 degrees of freedom, an adequate sample size. Eight animals per treatment group gives 24 degrees of freedom, also acceptable. Therefore, for brain metastasis assays, 8–10 mice were used in each group; exact numbers for each experimental series are included in the relevant figure legends. For drug treatment experiments, mice were inoculated with cancer cells and randomly assigned to treatment groups. Mice dying less than 24 h after tumour inoculation were excluded from analysis. Gap junction modulators and chemotherapeutic agents were blindly administered in the MSKCC Antitumour Assessment Core.

Immunohistochemical staining. Mouse brains were fixed with 4% paraformaldehyde, sectioned by vibratome (Leica) or cryostat (Leica) and stained following established protocols⁶. For brain slice assays⁶, 250- μ m thick slices of adult mouse brain were prepared with a vibratome (Leica) and placed on top of 0.8 μ m pore membranes (Millipore) in brain slice culture medium (DMEM, complete HBSS, 5% FBS, 1 mM L-glutamine, 100 IU ml⁻¹ penicillin, 100 μ g ml⁻¹ streptomycin). Approximately 3×10^5 cancer cells were placed on the surface of the slice. After 48 h of incubation, brain slices were fixed with 4% paraformaldehyde, and stained. For immunostaining in chamber slide cultures, cells were fixed with 4% paraformaldehyde and stained. Antibodies used for immunochemical staining are listed in Supplementary Information. Images were acquired with Zeiss Axio Imager Z1 microscope or Leica SP5 upright confocal microscope, and analysed with ImageJ, Imapis and Metamorph softwares. Antibodies used for immunostaining are listed in Supplementary Information.

Knockdown and overexpression constructs. For stable knockdown of Cx43, PCDH7, cGAS and STING, we used shRNAs in GiPZ lentiviral vector. For inducible knockdown, shRNAs in TRIPZ lentiviral vector were used. 1 μ g ml⁻¹ doxycycline hyclate (Sigma-Aldrich) was added to induce the expression of shRNA. Targeted sequences of shRNAs are listed in the Supplementary Information. pBabe-Puro-IK-Balpa-mut (Addgene) was used for stable expression of SR-IKBo. For expression of wild-type Cx43 (Origene) or Cx43(T154A) mutant (ACC to GCC), we used the pLVX vector.

mRNA and protein detection. Total RNA was extracted using the PrepEase RNA spin kit (USB). To prepare cDNA, 1 μ g of total RNA was treated using the

Transcriptor First Strand cDNA synthesis kit (Roche). CX43, CX30 and CX26 expression were quantified by Taqman gene expression assay primers: (CX43: Hs00748445_s1, Mm00439105_m1; CX30: Hs00922742_s1, Mm00433661_s1; CX26: Hs00269615_s1, Mm00433643_s1; Applied Biosystems). Relative gene expression was normalized to expression of B2M (β 2-microglobulin; Hs99999907_m1, Mm00437762_m1). The PCDH7 primer pair was designed to detect all PCDH7 isoforms: 5'-AGTTCAACGTGGTCATCGTG-3' (sense), 5'-ACAATCAGGAGTTGTTGCTC-3' (antisense). Reactions were performed using SYBR Green I Master Mix (Applied Biosystems). Quantitative expression data were analysed using an ABI Prism 7900HT Sequence Detection System (Applied Biosystems). For western immunoblotting, cell pellets were lysed with RIPA buffer and protein concentrations determined by BCA Protein Assay Kit (Pierce). Protein lysates of primary human astrocytes, microglia and HBMECs were purchased from ScienCell. Proteins were separated by SDS-PAGE and transferred to nitrocellulose membranes (BioRad). Antibodies used for western blotting are listed in Supplementary Table 1.

Dye transfer and EdU transfer assays. Monolayers of cancer cells or astrocytes were labelled with 2.5 μ g ml⁻¹ calcein Red-Orange AM dye (Life Technologies) at 37 °C for 30 min. Single-cell suspensions were mixed at a ratio of 2:1 labelled: unlabelled cells for 6 h. Certain experiments used a mix of three cell populations, MDA231-BrM2 (GFP⁺), HBMECs (pre-labelled with Cell Proliferating Dye Fluor@670, eBioscience), and unlabelled astrocytes. Dye transfer was visualized by Zeiss LSM 5 Live confocal microscopy (20-min time-lapse) or quantified by FACScalibur flow cytometry (BD Biosciences).

Cancer cell and astrocyte co-culture experiments. Astrocytes and cancer cells were mixed at ratio of 1:1. For apoptosis assays, overnight co-cultures were treated with 500 ng ml⁻¹ sFasL (Peprotech) in serum-free media, 500 nM carboplatin (Sigma-Aldrich) or 25 nM paclitaxel (Sigma-Aldrich) for 24 h. Single-cell suspensions were stained with allophycocyanin-conjugated cleaved caspase 3 antibody (Cell Signaling), apoptotic GFP⁺ cancer cells were detected by flow cytometry. For conditioned media analysis, media were collected after 24 h, and cytokines in the conditioned media were either identified using Human Cytokine Array (R&D systems) or measured by human or mouse IFN α or TNF ELISA kits (R&D systems) as relevant. To detect the activity of IFN α or TNF in the collected conditioned media, cancer cells were treated with the collected conditioned media for 2 h and phosphorylation status of STAT1 or NF- κ B p65 was determined by western blotting. For cGAMP and TANK-binding kinase 1 (TBK1)–IRF3 activation experiments, cancer cells and astrocytes were co-cultured for 18 h. To stimulate astrocytes without co-culture, astrocytes were transfected with cGAMP (4 μ g ml⁻¹) with Lipofectamine 2000, according to the manufacturer's protocol, as previously described⁴³. The phosphorylation status of TBK1, IRF3 was determined by western immunoblotting. Nuclear translocation of IRF3 was determined by immunofluorescence staining with Zeiss LSM 5 Live confocal microscopy. cGAMP levels were determined by LC-MS/MS. Detailed methods for cGAMP detection and quantification are described in Supplementary Information.

TRAP. eGFP-L10a-expressing cancer cells were co-cultured with astrocytes for 24 h. Following previously published protocols^{24,44}, mRNA purified from cancer cells was used for library construction with TruSeq RNA Sample Prep Kit v2 (Illumina) following the manufacturer's instructions. Samples were barcoded and run on a HiSeq 2000 platform in a 50 bp/50 bp paired-end run, using the TruSeq SBS Kit v3 (Illumina). An average of 50 million paired reads were generated per sample.

Cytokine treatment and pathway reporter assays. Cancer cells were treated with 10 U ml⁻¹ (39 U ng⁻¹) recombinant IFN α (R&D Systems) or 10 pg ml⁻¹ recombinant TNF (R&D Systems) in combination with carboplatin or taxol (Sigma-Aldrich) for 24 h. Apoptosis was quantified by Caspase-Glo 3/7 assay (Promega). For NF- κ B reporter assays, the NF- κ B responsive sequence from the pHAGE NFKB-TA-LUC-UBC-dTomato-W construct (Addgene)⁴⁵ was cloned into a pGL4.82 *Renilla* luciferase reporter (Promega). Cancer cells were co-transfected with this vector and a LeGO-C2 mCherry vector (Addgene). *Renilla* luciferase activity was determined using RenillaGlo Luciferase system (Promega). Red fluorescence signal was used to normalize transfection efficiency.

Split luciferase assay. Fusion cDNAs were generated by deleting the stop codon in human CX43 (Origene), PCDH7 (Origene), CDH1 (E-cadherin; Addgene) or CDH2 (N-cadherin; Addgene) cDNAs and splicing the N-terminal or C-terminal fragment of firefly luciferase²² (Addgene). Constructs were cloned into pLVX lentiviral expression vector and transduced into non-GFP-luciferase-labelled parental MDA-MB-231 or H2030 cells. To detect luciferase activity, 7.5 mg ml⁻¹ D-luciferin potassium salt was added in the culture media. BLI was performed by IVIS Spectrum Xenogen instrument, using Living Image software v.2.50.

Cytosolic dsDNA detection. For visualization of dsDNA, cells were immunostained with anti-dsDNA antibody. Anti-GFP antibody staining was used to delineate cancer cell bodies, DAPI to distinguish nuclei, and anti-CoxIV antibody

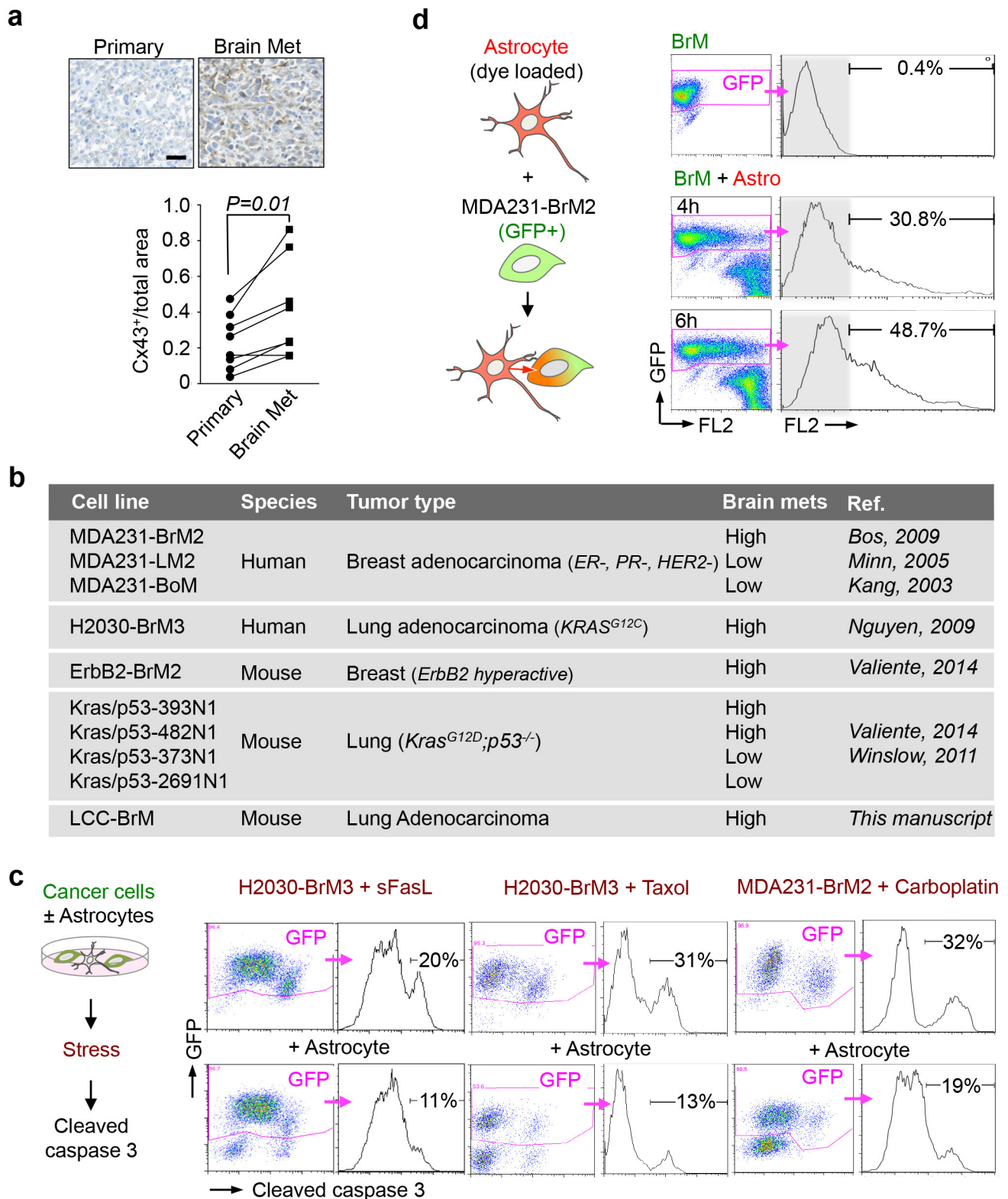
to distinguish mitochondria. Phalloidin staining (Molecular Probe) was used to delineate astrocyte cell bodies. For quantification of dsDNA, nuclear, cytosolic and mitochondrial fractions were prepared using a mitochondria isolation kit (Thermo Scientific). DNA from all subcellular fractions was purified by QIAamp DNA mini kit (Qiagen) and quantified by QuantoFluor dsDNA system (Promega).

Bioinformatic and statistical analysis. Bioinformatic analysis was performed in R (ver. 3.1.2) unless otherwise noted. The data were analysed using the TopHat2-HTSeq-DESeq2 pipeline^{46–48}. Differential gene expression was compared with cooksCutoff and independentFiltering turned off. Scatter plot showing fold changes was produced using the ggplot2 package. Principal component analysis was performed using prcomp. Pathway gene response signatures were analysed and scored by the sum of z-score method⁴⁴, as previously described^{12,49}. Multiple hypothesis testing was adjusted using the Benjamini and Hochberg false discovery rate method. Statistical analysis was performed using GraphPad software (Prism) and Student's *t*-test (two-tailed). *P* < 0.05 was considered statistically significant. Values are mean ± s.e.m.

Clinical sample analysis. All tissues were obtained from the MSKCC Department of Pathology in compliance with the MSKCC Institutional Review Board under Biospecimen Research Protocol 15-204. Informed consent was obtained from all subjects. *CX43* and *PCDH7* transcript levels were analysed in the microarray data of primary breast cancer (EMC-MSK) and adenocarcinoma data sets (MSKCC set2, GSE3141 and GSE8893). Multiple probes mapping to the same gene were combined by selecting the probe with maximal variance across samples. TNBC subtypes were identified based either on clinical annotation of the data set or on *ESR1* and *ERBB2* transcript levels. The hazard ratio of the *CX43* and *PCDH7* values was computed based on Cox proportional hazards model, as implemented by the 'coxph' command in R. *P* values were calculated from a Cox proportional hazard model, with *CX43* and *PCDH7* expression treated as a continuous variable.

For Cx43 immunohistochemistry, normal lung tissue array (75 cases), primary TNBC tissue array (98 cases) and primary NSCLC tissue array (138 cases) were purchased from US Biomax. Paraffin-embedded tissue microarrays from brain metastases (117 case of TNBC, 91 cases of NSCLC) were obtained from the MSKCC Department of Pathology. Immunohistochemical staining for Cx43 was performed by the MSKCC Pathology Core Facility using standardized, automated protocols. For matched primary-brain metastatic lesions, Cx43 staining was quantified by positive staining area (Metamorph software).

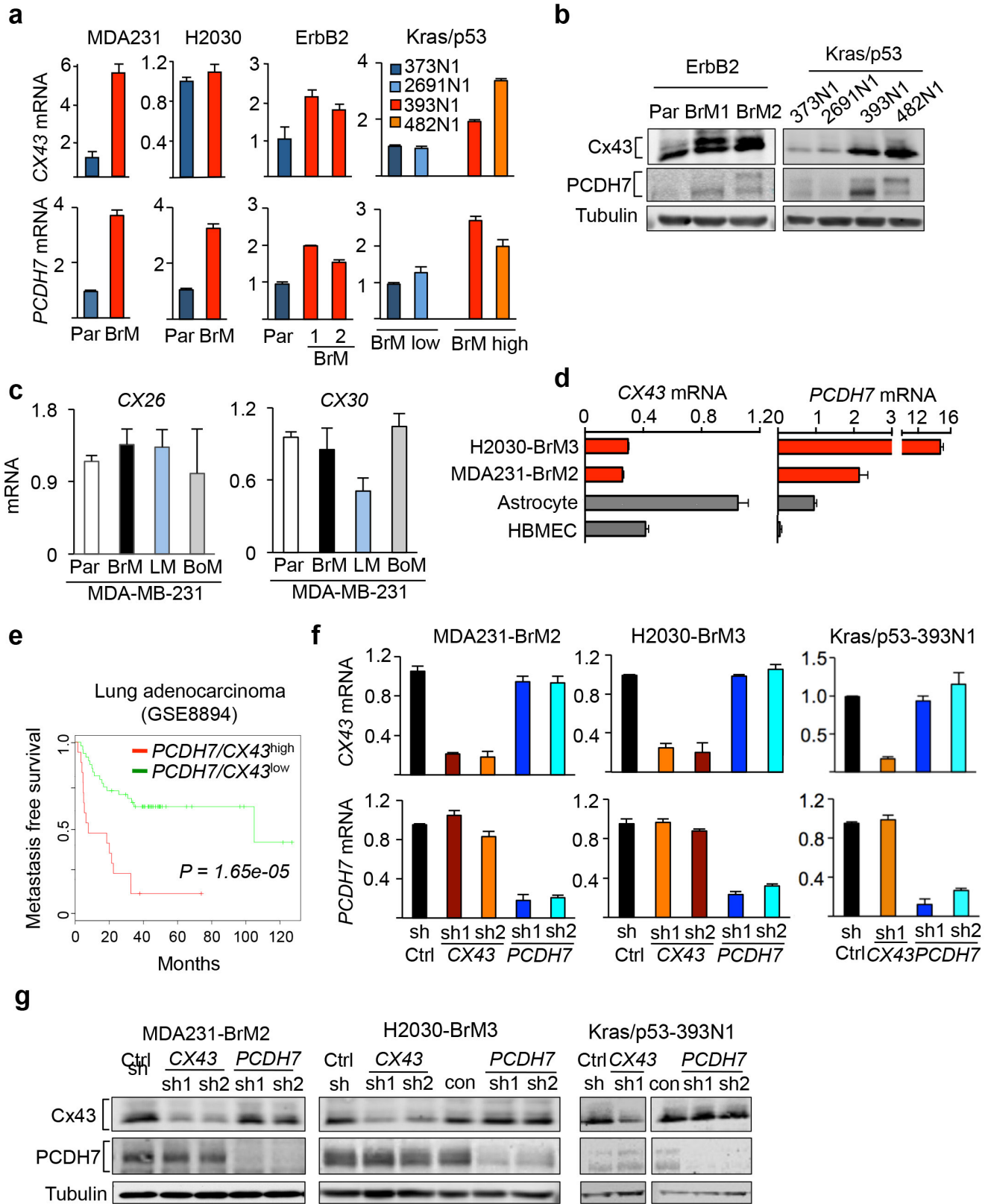
43. Sauer, J. D. *et al.* The *N*-ethyl-*N*-nitrosourea-induced *Goldenticket* mouse mutant reveals an essential function of *Sting* in the *in vivo* interferon response to *Listeria monocytogenes* and cyclic dinucleotides. *Infect. Immun.* **79**, 688–694 (2011).
44. Zhang, X. H. *et al.* Selection of bone metastasis seeds by mesenchymal signals in the primary tumor stroma. *Cell* **154**, 1060–1073 (2013).
45. Wilson, A. A. *et al.* Lentiviral delivery of RNAi for *in vivo* lineage-specific modulation of gene expression in mouse lung macrophages. *Mol. Ther.* **21**, 825–833 (2013).
46. Anders, S. *et al.* Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature Protocols* **8**, 1765–1786 (2013).
47. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
48. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
49. Gatz, M. L. *et al.* A pathway-based classification of human breast cancer. *Proc. Natl Acad. Sci. USA* **107**, 6994–6999 (2010).
50. Minn, A. J. *et al.* Genes that mediate breast cancer metastasis to lung. *Nature* **436**, 518–524 (2005).
51. Kang, Y. *et al.* A multigenic program mediating breast cancer metastasis to bone. *Cancer Cell* **3**, 537–549 (2003).



Extended Data Figure 1 | Cancer cell–astrocyte interactions.

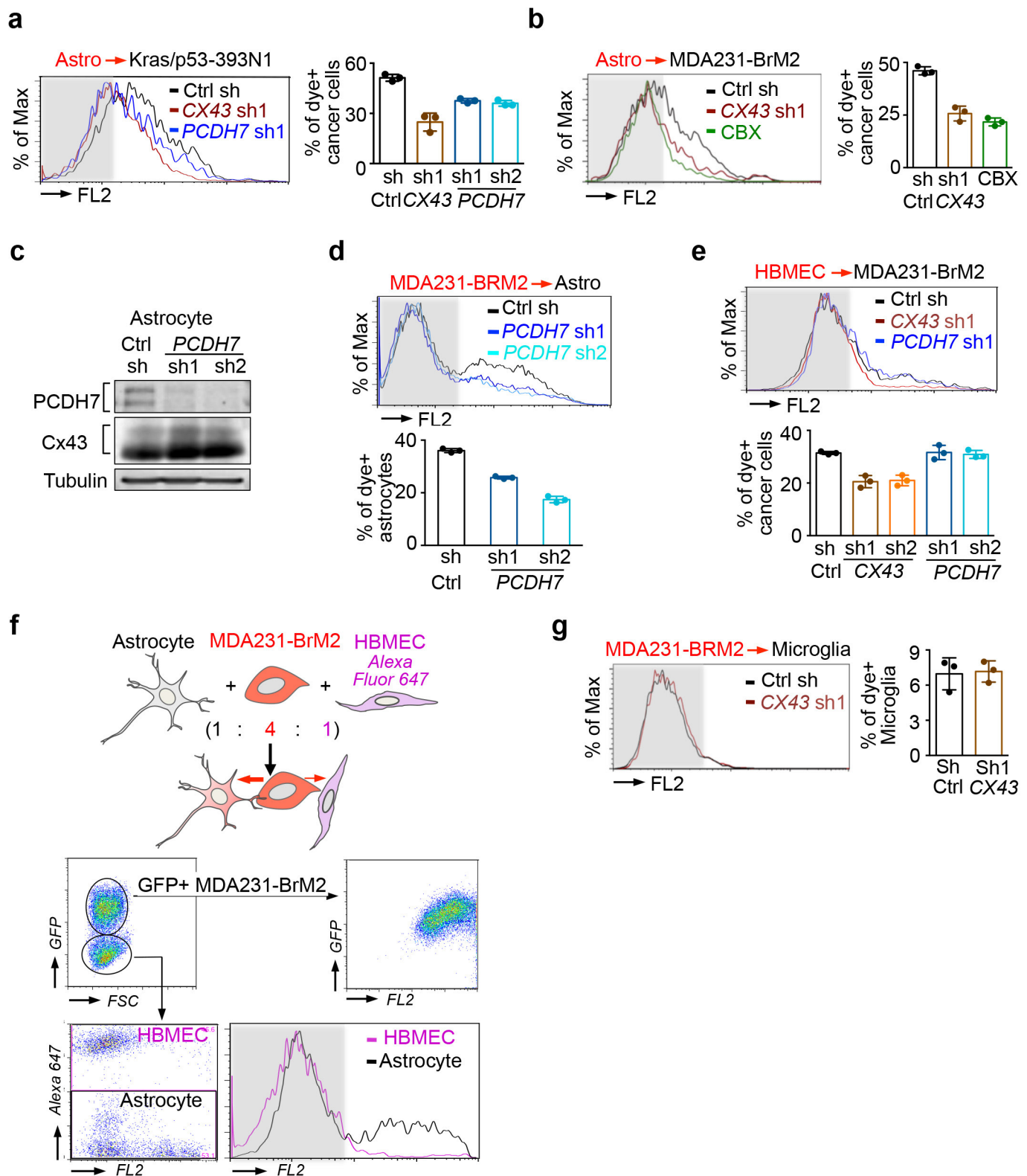
a, Representative images and quantification of Cx43 immunostaining in matched primary and brain metastatic samples from patients with NSCLC. Scale bar, 100 μ m ($n=8$ patients). **b**, Cancer cells used in this study. The following references are cited in the table: 3, 6, 12, 13, 50

and 51. **c**, Astrocyte co-culture protects cancer cells. As illustrated in schema (left), cleaved caspase 3⁺ GFP⁺ apoptotic BrM cells were quantified by flow cytometry after sFasL addition or chemo-treatments (3 independent experiments). **d**, Flow cytometric quantification of dye transfer from astrocytes to MDA231-BrM2 cells over time (3 independent experiments).



Extended Data Figure 2 | Increased expression of Cx43 and PCDH7 in brain metastatic cancer cells and astrocytes. **a**, CX43 and PCDH7 mRNA in parental and BrM cells. Data are mean \pm s.e.m. ($n = 3$ independent experiments in triplicate). **b**, CX43 and PCDH7 western blotting in ErbB2 parental and brain metastatic cells, as well as Kras/p53 cell lines ($n = 3$ independent experiments). **c**, CX26 and CX30 mRNA in MDA231 parental cell lines and the metastatic derivatives of brain (BrM2), lung (LM) and

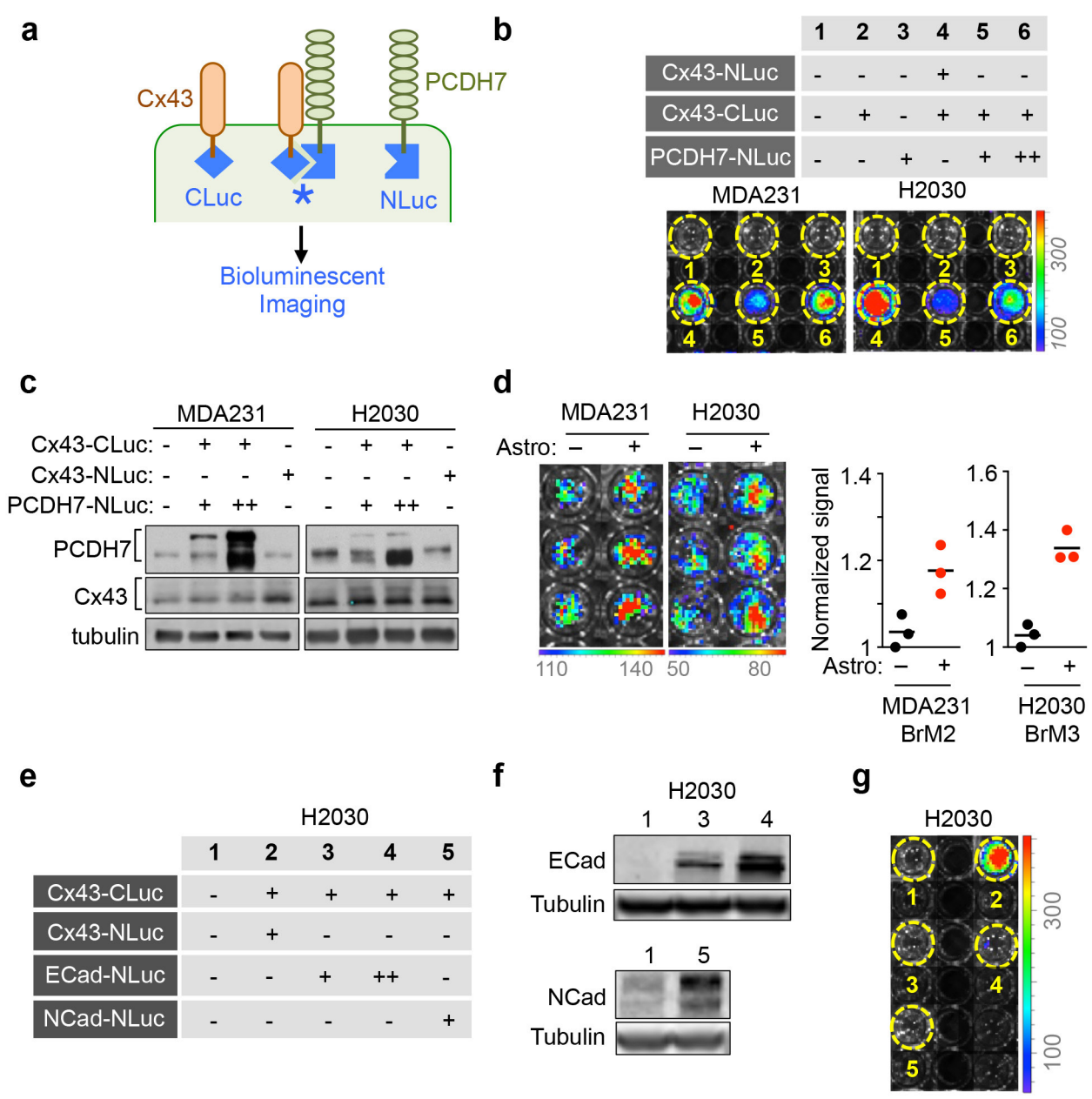
bone (BoM). **d**, CX43 and PCDH7 mRNA in BrM cells compared to brain cells ($n = 3$ independent experiments). **e**, Kaplan-Meier plot illustrates the probability of cumulative metastasis-free survival in 63 cases (GSE8893) of lung adenocarcinoma based on CX43 and PCDH7 expression in the primary tumour. **f**, **g**, Knockdown of Cx43 and PCDH7 with shRNAs as assessed by reverse transcriptase PCR (RT-PCR) (**f**) and western blotting (**g**). Data are mean \pm s.e.m. ($n = 3$ independent experiments in triplicate).



Extended Data Figure 3 | PCDH7 facilitates gap junction communication.

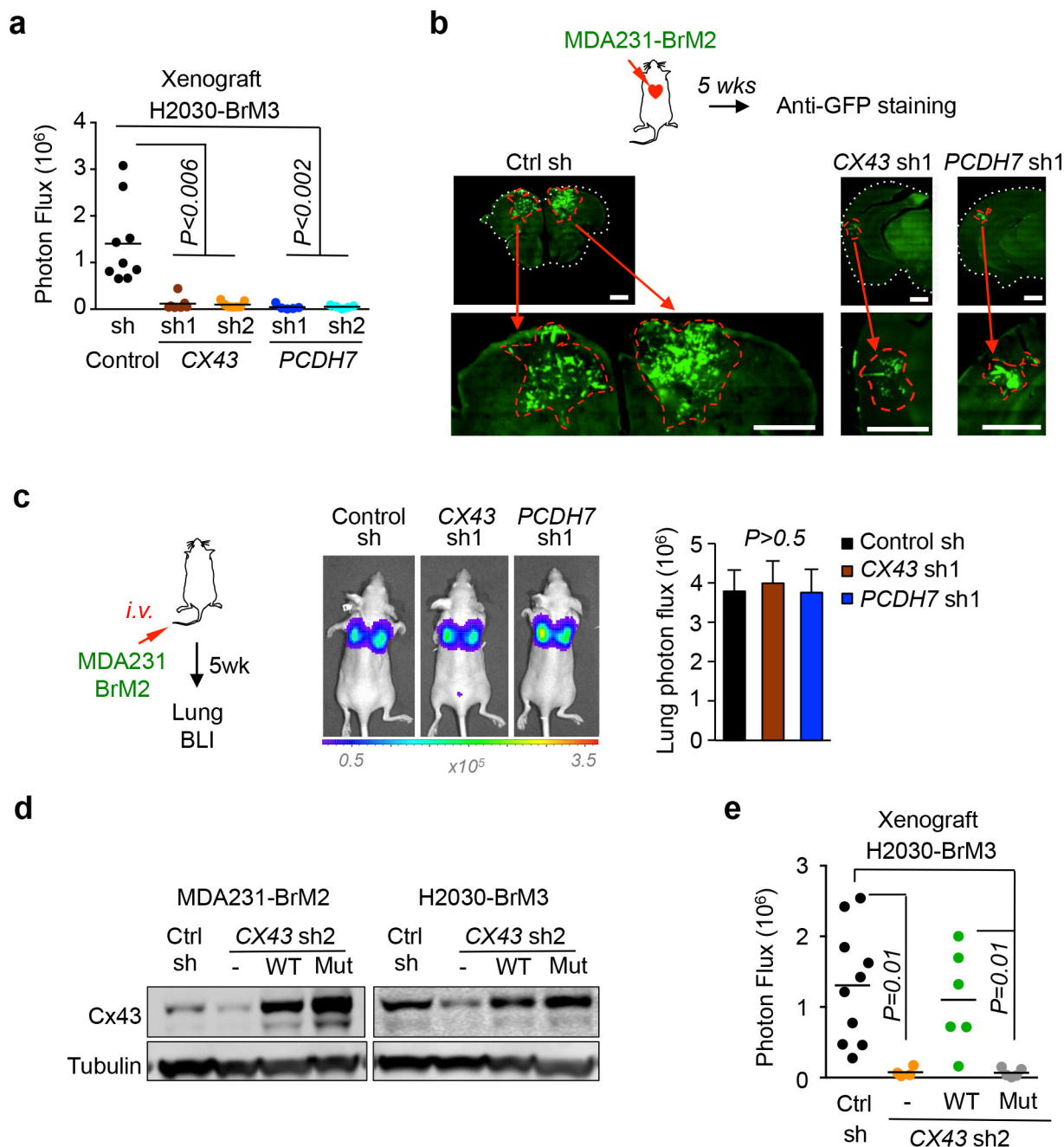
a, b, Histograms and quantification of dye transfer from astrocytes to control and Cx43- or PCDH7-depleted Kras/p53-393N1 cells (**a**), and from astrocytes to control or Cx43-depleted MDA231-BrM2 cells, in comparison to carbenoxolone (50 μ M) treatment (**b**). **c, d**, PCDH7 in astrocytes facilitates gap junctions. **c**, PCDH7 immunoblotting of control or PCDH7-depleted astrocytes. **d**, Quantification of dye transfer from

MDA231-BrM2 cells to PCDH7-depleted astrocytes (**d**). **e**, Quantification of dye transfer from HBMECs to control, Cx43- or PCDH7-depleted MDA231-BrM2 cells. **f**, Dye transfer from MDA231-BrM2 cells to a mixed population of astrocytes and HBMECs. **g**, Quantification of dye transfer from control or Cx43-depleted MDA231-BrM2 cells to human microglia. For dye transfer assays, values are mean \pm s.e.m. ($n \geq 2$ independent experiments in triplicate).



Extended Data Figure 4 | Cx43 directly interacts with PCDH7, but not with E-cadherin or N-cadherin. **a**, Schema illustrating split luciferase assay. Fusion constructs of PCDH7 and Cx43 were created with either NLuc or CLuc. When these proteins are brought into proximity, luciferase is functionally reconstituted, producing photons of light. **b**, Cx43 and PCDH7 constructs fused with NLuc and CLuc were expressed in parental cell lines. The table (top) numerically identifies the cell line combinations used in the assays (bottom), and BLI of a representative plate. **c**, Cx43 and PCDH7 western immunoblotting in cancer cells overexpressing

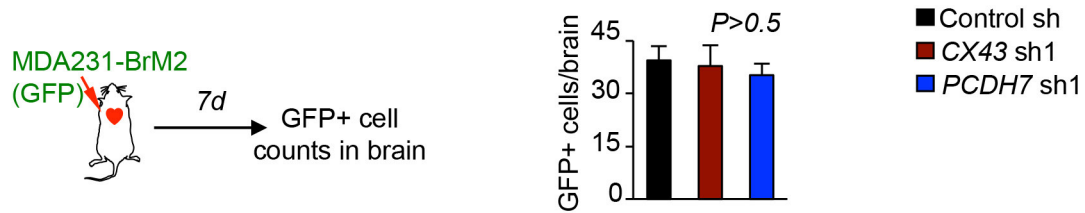
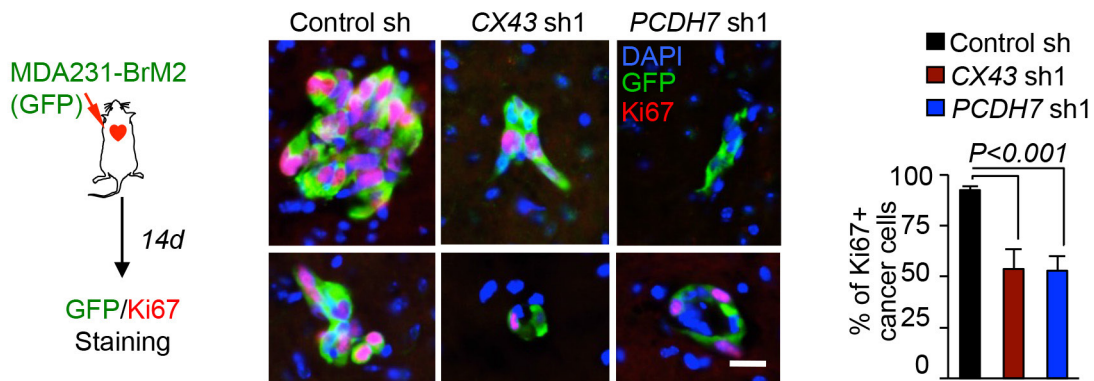
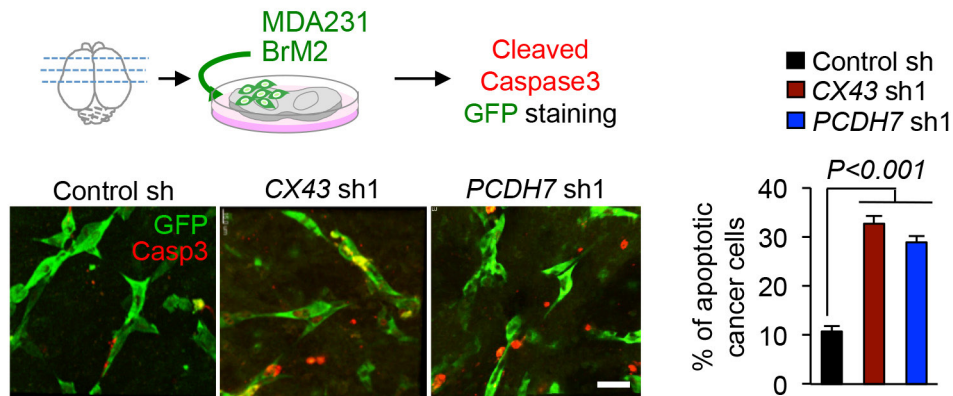
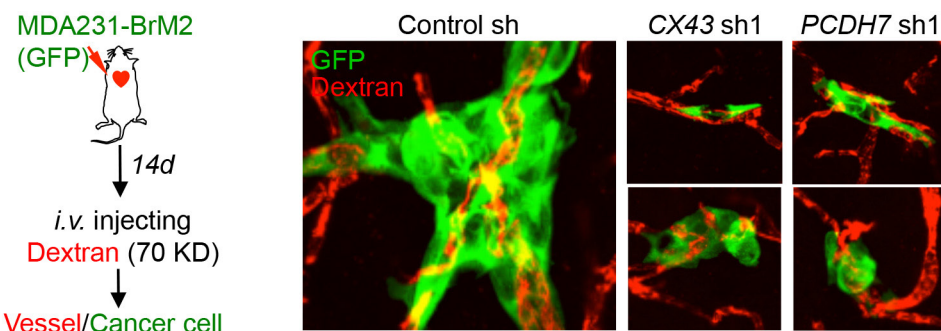
fusion proteins. **d**, Quantification of BLI after co-culture of Cx43-CLuc and PCDH7-NLuc cancer cells and astrocytes for 15 min (3 independent experiments) **e–g**, Luciferase split assay to detect Cx43–E-cadherin or Cx43–N-cadherin interactions. Cell line combinations used in the assays are numerically identified in the table (**e**), and confirmed by western immunoblotting (**f**). **g**, BLI of a representative assay plate; cell line combinations are indicated numerically ($n \geq 2$ independent experiments in **e–g**).



Extended Data Figure 5 | Inhibition of gap junction activity prevents brain metastatic outgrowth. **a**, BLI quantification of brain metastatic lesions formed by control, Cx43- or PCDH7-depleted H2030-BrM3 cells ($n = 2$ independent experiments with 9 mice total per group).

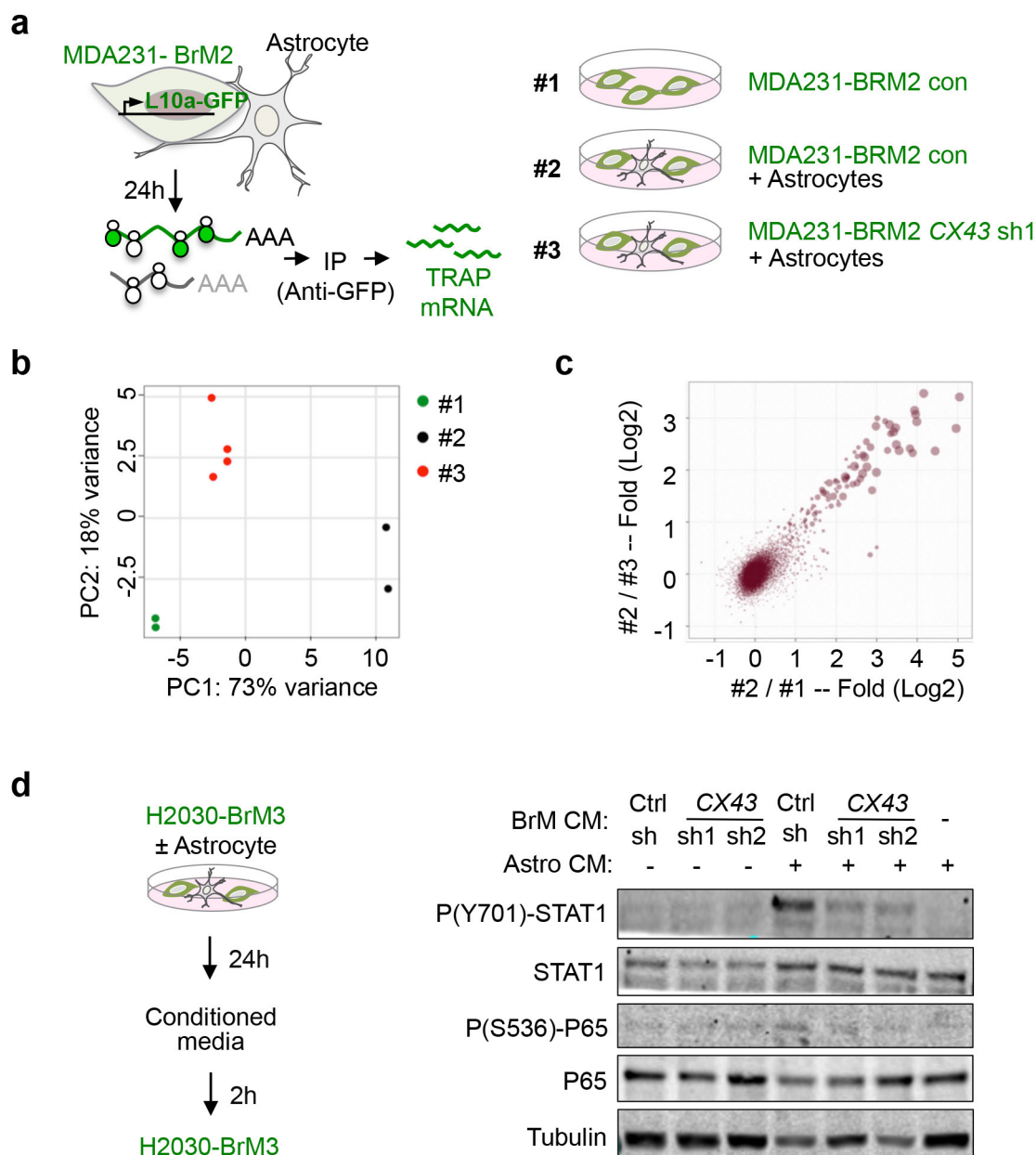
b, Representative images of GFP⁺ brain metastatic lesions formed by control, Cx43- or PCDH7-depleted MDA231-BrM2 cells. Brain sections or brain metastatic lesions are delineated by dotted white or red lines, respectively. Scale bars, 1,000 μm . **c**, BLI (images) and quantification

(bar graph) of lung metastatic lesions formed by MDA231-BrM2 cells. Values are mean \pm s.e.m. ($n = 2$ independent experiments with 5 mice total in each group). **d**, **e**, Gap-junction-mediated brain metastasis requires channel function of Cx43. Wild-type or T154A mutant Cx43 was re-expressed in Cx43-depleted (CX43 sh2) MDA231-BrM2 cells. Cx43 expression was detected by western blotting (**d**), and brain metastasis formed by these cells was quantified by BLI (**e**) ($n = 2$ independent experiments with 10 mice total per group).

a**b****c****d**

Extended Data Figure 6 | Cx43 and PCDH7 do not mediate early events of extravasation and vascular co-option in brain metastasis. **a**, Cx43 and PCDH7 do not mediate trans-BBB migration. Quantification of control, Cx43- or PCDH7-depleted MDA231-BrM2 cells in 7-day brain lesions was carried out as follows: at the indicated time point, mice were euthanized, brains were sectioned, 10% of the sections were immunostained, and all GFP⁺ cells in these sections were counted. Data are mean \pm s.e.m. ($n = 5$ brains in each group). **b**, Cx43 and PCDH7 mediate cancer cell colonization in 14-day brain lesions. Sectioning and staining were carried out as described in **a**. Representative images are GFP (green) and Ki67 (red) staining. DAPI, nuclear staining. Scale bar, 20 μ m. Bar graph is the proportion of Ki67⁺ cancer cells. Data are mean \pm s.e.m. ($n = 5$ brains in each group). **c**, Cx43 and PCDH7 mediate cancer cell survival.

MDA231-BrM2 cells expressing CX43 shRNA, PCDH7 shRNA or control shRNA were deposited onto living brain sections, five brain slices were seeded with cancer cells of each type. After 48 h, slices were fixed and stained for GFP (green) and cleaved caspase 3 (Casp3) (red) staining. Representative images are shown. Scale bar, 30 μ m. After staining, all GFP⁺ cells were counted on each slice. GFP⁺ cells with caspase 3⁺ staining were scored as 'apoptotic'. Histogram shows proportion of caspase 3⁺ apoptotic cancer cells. Data are mean \pm s.e.m. ($n = 5$ brain slices in each group). **d**, Cx43 and PCDH7 do not affect vascular co-option of cancer cells in 14-day brain lesions. Representative images are GFP (green) staining and vascular structure filled with TRITC dextran (red). Scale bar, 20 μ m ($n = 2$ independent experiments).



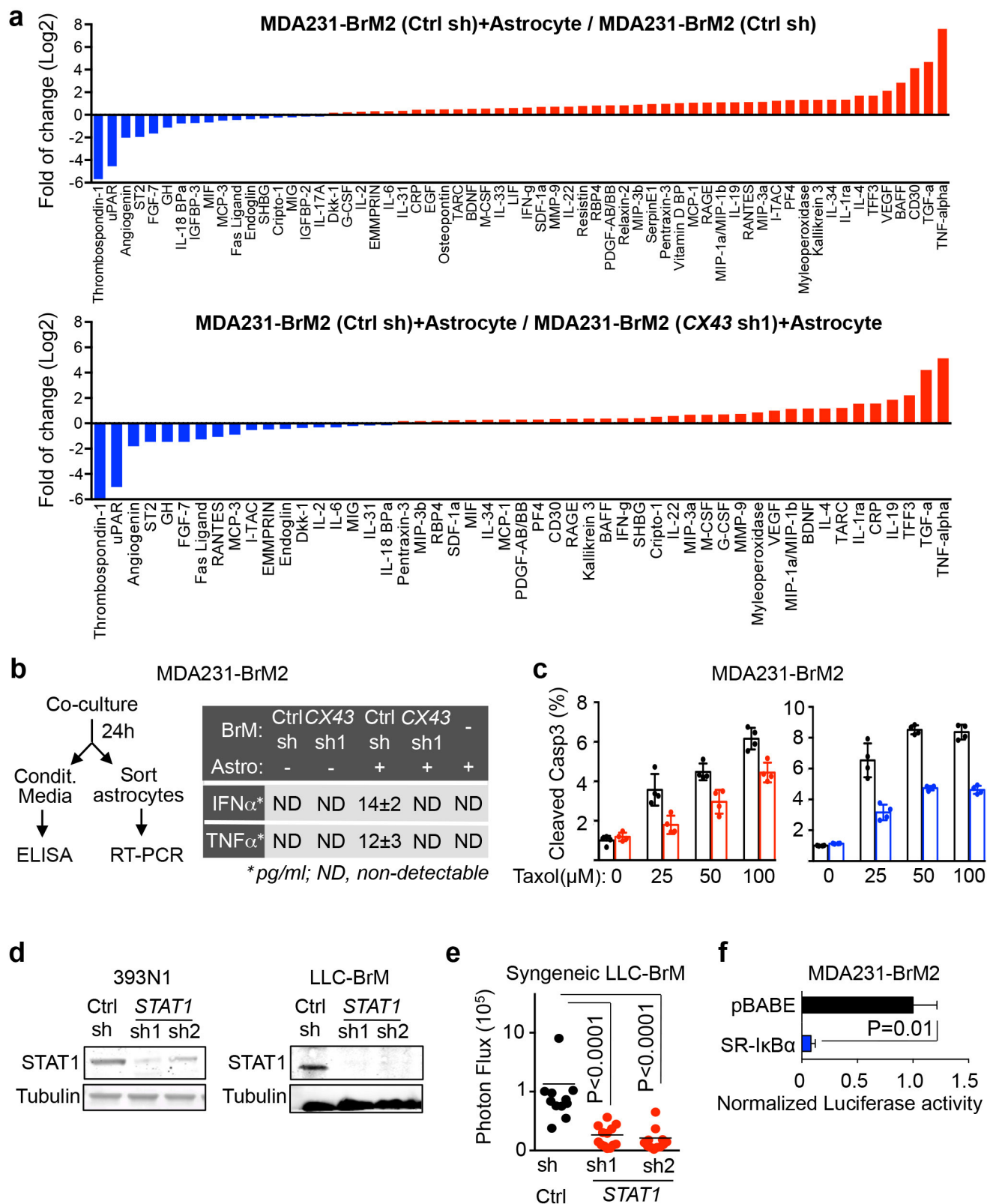
Extended Data Figure 7 | TRAP after cancer cell astrocyte co-culture.

a, Schematic illustration of TRAP experimental set up to isolate translating mRNA from MDA231-BrM2 cells under three conditions (1, 2 and 3).

b, Principle component (PC) analysis of TRAP mRNA sequencing.

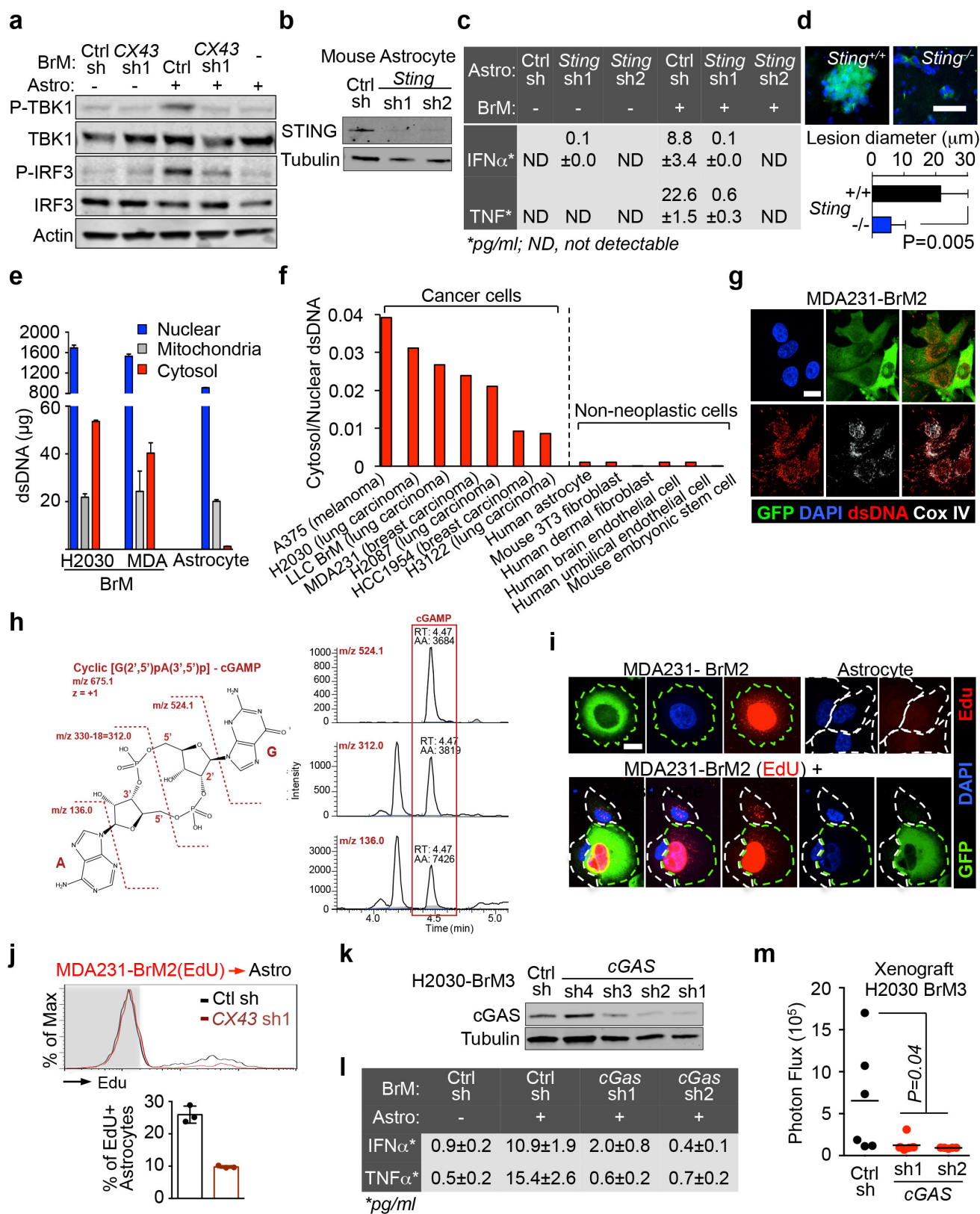
c, Scatter plot of log₂ fold changes regulated by astrocytes and gap junction

communications between BrM cells and astrocytes. **d**, STAT1 and NF- κ B p65 phosphorylation in H2030-BrM3 cells after a 2 h incubation with conditioned media from astrocyte co-cultures. Conditioned media samples were collected after 24 h co-culture of astrocytes with control or Cx43-depleted H2030-BrM3 cells ($n = 3$ independent experiments).



Extended Data Figure 8 | Gap-junction-generated signalling activates IFN and NF- κ B pathways in cancer cells. **a**, Cytokine array analysis of conditioned media collected after 24h co-culture of human astrocytes with control or Cx43-depleted MDA231-BrM2 cells. The \log_2 fold changes are plotted. **b**, Schematic of co-culture conditioned media collection and human astrocyte re-isolation (left) ELISA of IFN α and TNF in conditioned media from astrocyte co-cultures with the indicated MDA231-BrM2 cells (right) Data are mean \pm s.e.m. ($n \geq 2$ independent experiments with 4 total replicates). **c**, Relative levels of cleaved caspase 3 in MDA231-BrM2 cells treated with various concentrations of carboplatin

in the presence or absence of 10 U ml $^{-1}$ (39 U ng $^{-1}$) IFN α or 10 pg ml $^{-1}$ TNF. Data are mean \pm s.e.m. ($n = 5$ technical replicates over 3 independent experiments). **d**, STAT1 levels in control and STAT1-knockdown LLC-BrM and 393N1 cells. **e**, Quantification of BLI signal from brain metastases formed by syngeneic LLC-BrM control, or STAT1-knockdown cells ($n = 2$ independent experiments with 12–15 mice total per group). **f**, NF- κ B *Renilla* luciferase reporter assay in MDA231-BrM cells expressing control pBABE or SR-IkBa vector. Data are mean \pm s.e.m. ($n = 3$ technical replicates).

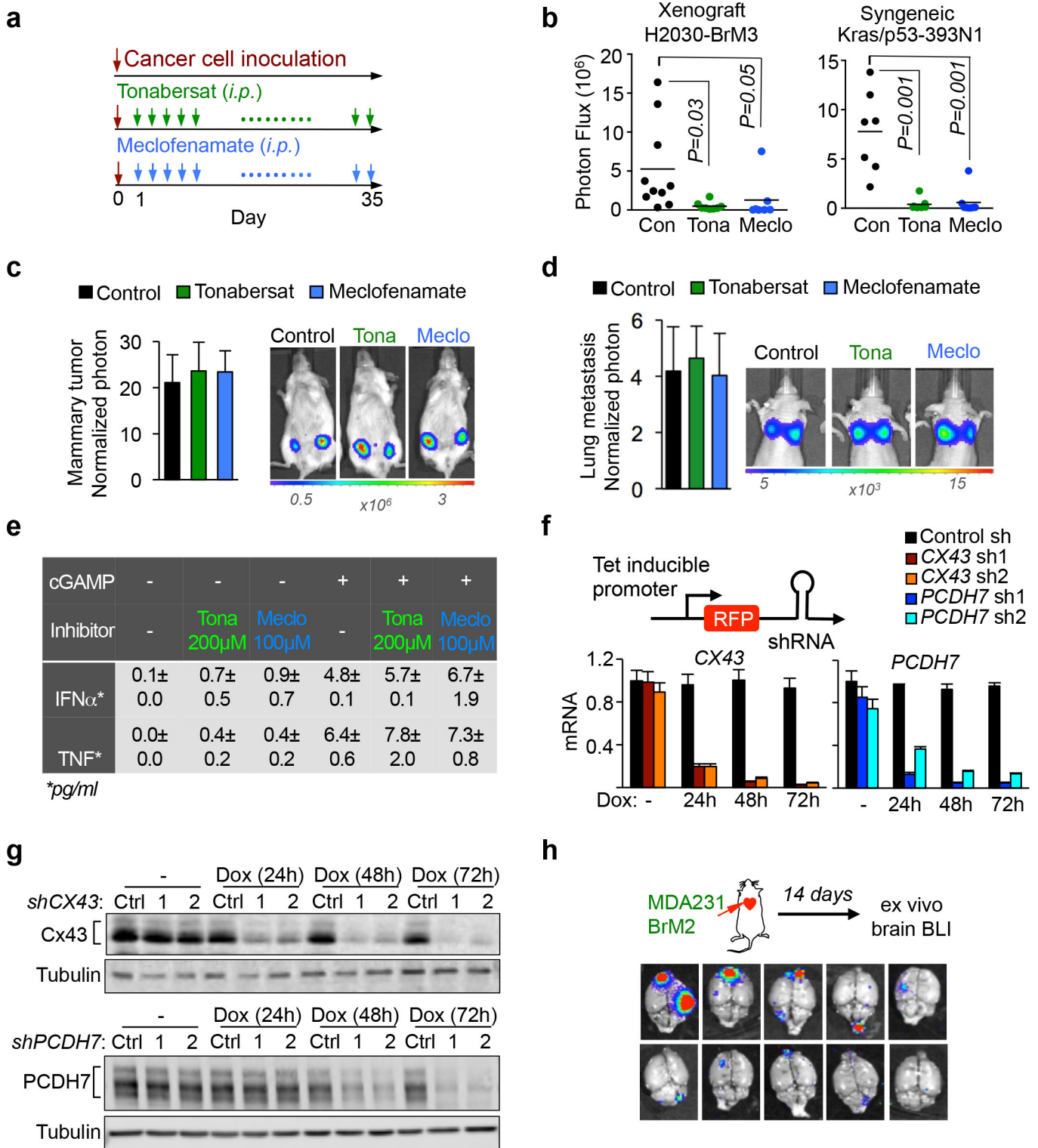


Extended Data Figure 9 | See next page for caption.

Extended Data Figure 9 | Gap junctions initiate cytosolic DNA response in astrocytes.

a, Control or Cx43-depleted H2030-BrM3 cells were co-cultured for 18 h with/without astrocytes, and subjected to immunoblotting analysis of phosphorylated TBK1 and IRF3 ($n = 3$ independent experiments). **b**, Immunoblot of mouse astrocytes depleted of STING with control (non-silencing) or *Sting* shRNAs. **c**, Mouse IFN α and TNF were quantified in the conditioned medium after co-culture by ELISA ($n = 2$ independent experiments with 3 replicates each). **d**, LLC-BrM growth in syngeneic C57Bl6 mice hosts wild-type (+/+) or knockout (−/−) for *Sting*. Bottom, diameter of brain metastases. Scale bar, 50 μ m. Brains from all mice ($n = 22$) were sectioned, immunostained, and measured. All GFP⁺ brain metastases were quantified (2.8 ± 0.67 metastases per *Sting*^{+/+} mouse; 1.6 ± 0.55 in *Sting*^{−/−} mice). **e**, Quantification of dsDNA in the indicated cellular fractions from 2×10^7 H2030-BrM3, MDA231-BrM2 or human astrocyte cells. Data are mean \pm s.e.m. ($n = 3$ biological replicates; 2 independent experiments). **f**, Ratio of cytosolic dsDNA and nuclear dsDNA in indicated cancer cells

and non-neoplastic cells. **g**, Representative image of immunofluorescent staining of dsDNA, GFP and CoxIV (a mitochondrial marker) in MDA231-BrM2 cells. **h**, cGAMP identification. The peak at 4.47 min contains all three selected reaction monitoring (SRM) transitions specific for cGAMP. AA, automatically integrated peak area; RT, retention time. **i**, **j**, EdU-labelled MDA231-BrM2 cells were co-cultured with astrocytes for 6 h. Transfer of EdU-labelled DNA from cancer cells to astrocytes was visualized using confocal microscopy (**i**), or quantified by flow cytometry (**j**). **k**, Immunoblot of H2030-BrM3 cells depleted of cGAS with shRNAs or control shRNA. **l**, Human astrocytes, were cultured for 18 h with/without H2030-BrM cells expressing control or cGAS shRNA. Human IFN α and TNF were quantified in the conditioned medium by ELISA ($n = 2$ independent experiments in triplicate). **m**, Quantification of BLI signal from brain metastases formed by H2030-BrM3 cells depleted of cGAS with two independent shRNAs ($n = 2$ independent experiments with 6 mice total per group).



Extended Data Figure 10 | Inhibition of gap junction activity prevents brain metastatic outgrowth. **a–d**, After treatment with tonabersat or meclofenamate (**a**), brain metastasis (**b**), primary tumour growth in mammary fat pads (**c**), or lung metastasis (**d**) was quantified by BLI. Data are mean \pm s.e.m. ($n = 2$ independent experiments with 10 mice total in each group). **e**, Human astrocytes were treated with 200 μ M tonabersat or 100 μ M meclofenamate for 12 h before transfection with

cGAMP (4 μ g ml⁻¹) using Lipofectamine 2000 or Lipofectamine alone. Conditioned media was collected 18 h later and assayed for human TNF and IFN α by ELISA ($n = 2$ biological replicates). **f, g**, Knockdown of CX43 and PCDH7 in MDA231-BrM2 cells with tet-on inducible shRNA as assessed by RT-PCR (**f**) and western blotting (**g**), after doxycycline treatment *in vitro* ($n = 2$ independent experiments). **h**, Brain *ex vivo* BLI 14 days after inoculation of MDA231-BrM2 cells ($n = 10$ mice).

Synchronized mitochondrial and cytosolic translation programs

Mary T. Couvillion¹, Iliana C. Soto¹, Gergana Shipkovenska¹ & L. Stirling Churchman¹

Oxidative phosphorylation (OXPHOS) is a vital process for energy generation, and is carried out by complexes within the mitochondria. OXPHOS complexes pose a unique challenge for cells because their subunits are encoded on both the nuclear and the mitochondrial genomes. Genomic approaches designed to study nuclear/cytosolic and bacterial gene expression have not been broadly applied to mitochondria, so the co-regulation of OXPHOS genes remains largely unexplored. Here we monitor mitochondrial and nuclear gene expression in *Saccharomyces cerevisiae* during mitochondrial biogenesis, when OXPHOS complexes are synthesized. We show that nuclear- and mitochondrial-encoded OXPHOS transcript levels do not increase concordantly. Instead, mitochondrial and cytosolic translation are rapidly, dynamically and synchronously regulated. Furthermore, cytosolic translation processes control mitochondrial translation unidirectionally. Thus, the nuclear genome coordinates mitochondrial and cytosolic translation to orchestrate the timely synthesis of OXPHOS complexes, representing an unappreciated regulatory layer shaping the mitochondrial proteome. Our whole-cell genomic profiling approach establishes a foundation for studies of global gene regulation in mitochondria.

The large majority of cellular energy is produced by oxidative phosphorylation (OXPHOS) complexes within the mitochondrial inner membrane. These complexes consist of a mix of mitochondrial- and nuclear-encoded subunits, requiring the cell to coordinate completely separate gene expression machineries in order to match subunit expression with environmental demands for energy. The mitochondrial gene expression machinery is distinct from its nuclear/cytosolic counterparts, and has also diverged markedly from its bacterial correlates. Transcription is carried out by a single-subunit phage-related RNA polymerase¹ and translation by a dedicated ribosome (the mitoribosome) that is protein-rich compared to cytosolic and bacterial ribosomes². Mitochondrial transcripts are polycistronic and mRNAs have neither 5' caps nor Shine–Dalgarno sequences. In some species, including *S. cerevisiae*, poly(A) tails are also absent³. Mitochondria use modified genetic codes, deciphered by mitochondrial-encoded tRNAs⁴. Most notably, no mitochondrial gene-specific transcription factors have been characterized; instead, mitochondria contain mRNA-specific translational activators, generally present in limiting quantities, that have roles in initiation and/or elongation and, in some cases, in feedback control of OXPHOS complex assembly on subunit translation^{5–8}. Thus, the nuclear and mitochondrial genes are expressed by distinct machineries and controlled by disparate regulatory mechanisms. It remains unclear whether these radically different genomes coordinate their gene expression programs during any physiological response when OXPHOS synthesis is required, such as during mitochondrial biogenesis.

OXPHOS mRNAs are not coordinately induced

To analyse OXPHOS expression comprehensively, we used a set of quantitative approaches to monitor the levels and translation of mitochondrial- and nuclear-encoded RNA (Fig. 1a). To induce OXPHOS synthesis, we rapidly shifted *S. cerevisiae* cells from the fermentable carbon source glucose to non-fermentable glycerol, requiring reprogramming of gene expression to adapt to respiratory metabolism^{9,10} (Fig. 1b). As expected, steady-state protein levels of both mitochondrial- and nuclear-encoded OXPHOS subunits are induced as cells adapt to respiratory metabolism, and accumulate to high levels in cells undergoing log phase growth in glycerol (Extended Data Fig. 1). Mitochondrial

transcripts accumulate in response to the shift^{11,12}, as do nuclear-encoded OXPHOS mRNAs^{13,14}, but whether the transcript abundances rise concordantly is not clear. To quantify levels of both nuclear- and mitochondrial-encoded mRNAs we used rRNA depletion, as poly(A) selection would not capture mitochondrial messages, and included spike-in standards to allow quantification across samples. As is observed in most transcriptional programs, nuclear-encoded protein complex components are co-regulated at the RNA level¹⁵ (Extended Data Fig. 2a; full data set provided in Supplementary Table 1). The mitochondrial genome encodes eight major proteins that contribute to dual-origin complexes: the mitoribosome and OXPHOS complexes III–V. The genome also produces low levels of maturases that are required to process *COB* and *COX1* mRNAs (Extended Data Fig. 2b). The nuclear- and mitochondrial-encoded RNAs of the mitoribosome do not change appreciably across the time series, and so by default display similar dynamics (Extended Data Fig. 2c). By contrast, the levels of the nuclear- and mitochondrial-encoded OXPHOS complex RNAs increase during adaptation, but not coordinately (Fig. 1c). Whereas nuclear OXPHOS messages are induced rapidly in response to nutrient shift, mitochondrial OXPHOS messages are induced much more slowly. The difference in induction kinetics may reflect the absence of environment-responsive mitochondrial transcription factors.

Mitochondrial translation is dynamically regulated

Traditionally, mitochondrial translation has been monitored using metabolic labelling after inhibition of cytosolic translation by cycloheximide¹⁶, but this method requires specific buffer conditions and has poor time resolution. Thus, despite the existence of translational activators, it is not known whether translation of mitochondrial mRNAs is differentially regulated under normal physiological conditions, nor whether mitochondrial translation, like cytosolic translation¹⁷, responds rapidly to environmental changes. To quantitatively monitor mitochondrial translation under any growth condition with high time resolution, we re-engineered the ribosome profiling approach originally developed for cytosolic ribosomes¹⁸ through three major modifications. (1) Affinity purification by Flag-tagged mitoribosomal subunits replaced sucrose fractionation to separate 74S mitoribosomes from 80S cytosolic

¹Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA.

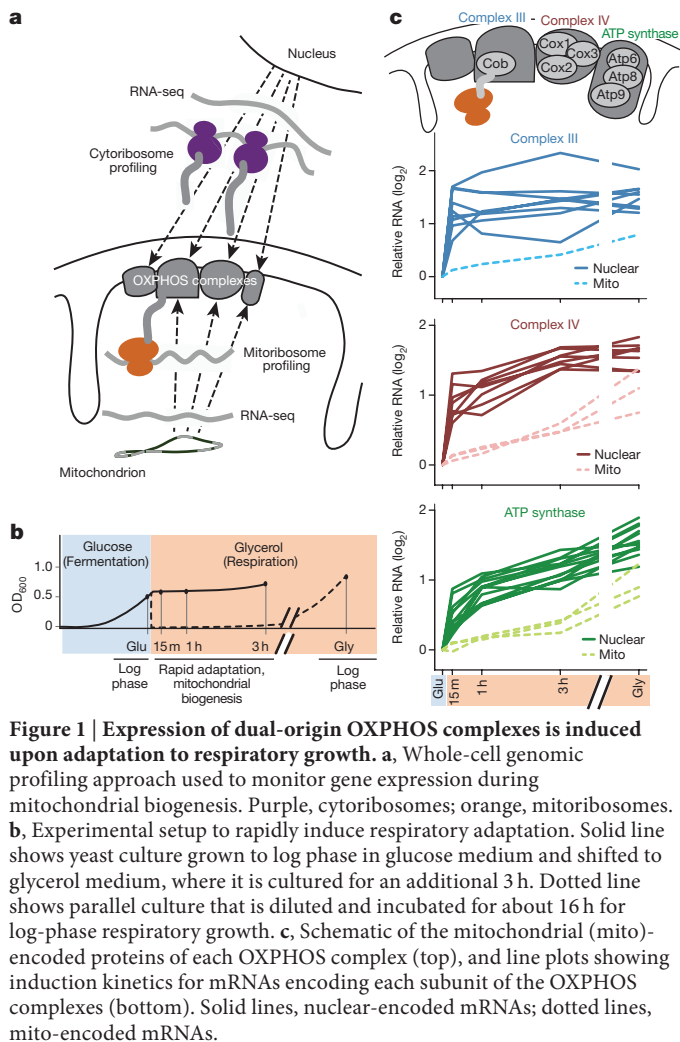


Figure 1 | Expression of dual-origin OXPHOS complexes is induced upon adaptation to respiratory growth. **a**, Whole-cell genomic profiling approach used to monitor gene expression during mitochondrial biogenesis. Purple, cytoribosomes; orange, mitoribosomes. **b**, Experimental setup to rapidly induce respiratory adaptation. Solid line shows yeast culture grown to log phase in glucose medium and shifted to glycerol medium, where it is cultured for an additional 3 h. Dotted line shows parallel culture that is diluted and incubated for about 16 h for log-phase respiratory growth. **c**, Schematic of the mitochondrial (mito)-encoded proteins of each OXPHOS complex (top), and line plots showing induction kinetics for mRNAs encoding each subunit of the OXPHOS complexes (bottom). Solid lines, nuclear-encoded mRNAs; dotted lines, mito-encoded mRNAs.

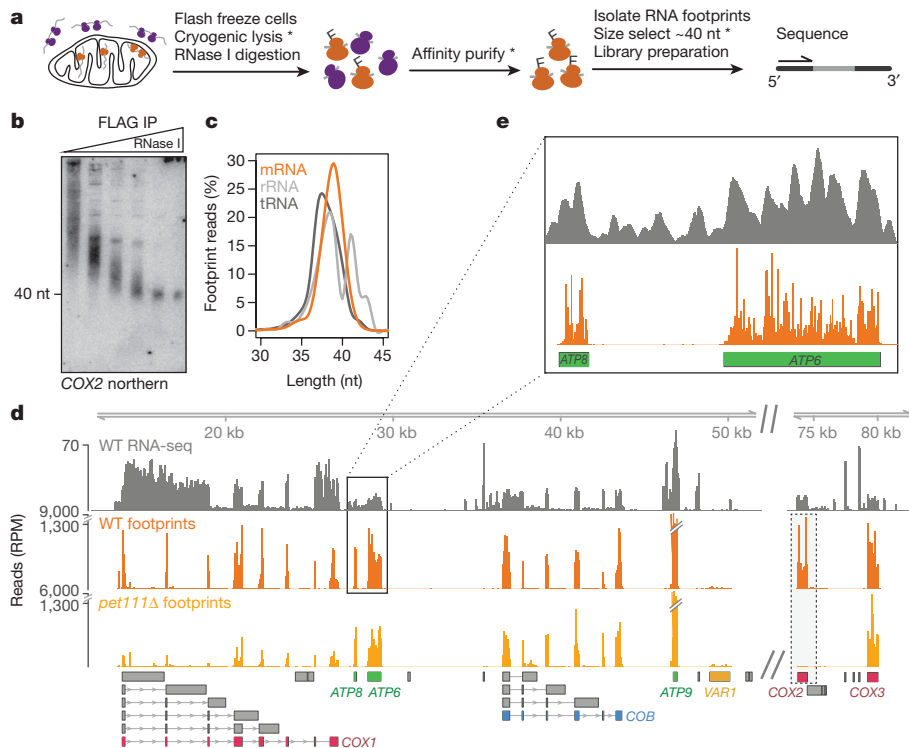


Figure 2 | Mitoribosome profiling provides a genome-wide readout of mitochondrial translation. **a**, Schematic of mitoribosome profiling protocol. Asterisks denote steps in which major modifications were required to capture mitoribosome footprints rather than cytoribosome footprints. **b**, RNase I titration (0, 50, 125, 250, 500, and 1,000 U ml⁻¹) followed by mitoribosome immunoprecipitation (IP) via MrpS17-Flag. For gel source data, see Supplementary Fig. 1. nt, nucleotides. **c**, Length distribution for mitoribosome profiling reads that map to mitochondrial-encoded mRNA in comparison to contaminating reads that map to rRNA or tRNA. **d**, Genome-wide view of mitochondrial open reading frames (ORFs) with mapped RNA-seq reads and mitoribosome profiling footprint reads (inferred A site). Lack of COX2-mapped reads in the *pet111Δ* strain is highlighted. Major ORFs are coloured. Grey annotations are maturase genes (note low level of translation) and tRNA genes. **e**, Expanded view of the region encoding the polycistronic ATP8/ATP6 transcript.

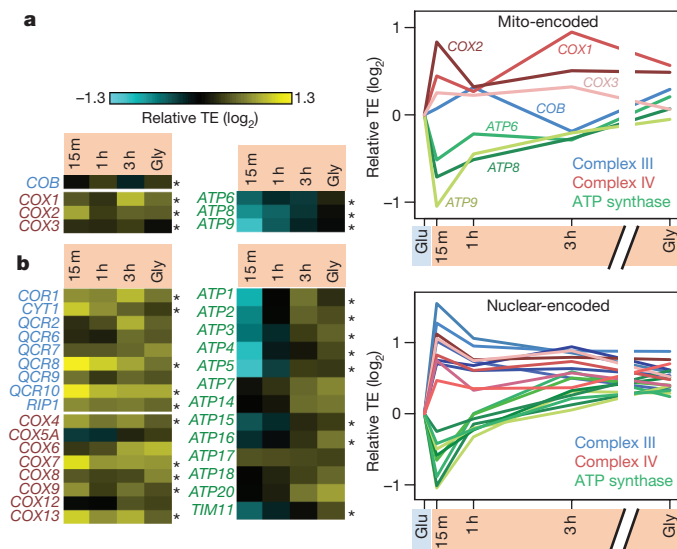


Figure 3 | Mitochondrial and cytosolic translation of OXPHOS mRNAs is rapidly and synchronously regulated. **a, b,** Fold changes in translation efficiency (TE) compared to log-phase glucose growth for the OXPHOS subunits synthesized in mitochondria (**a**; values are averages of two experiments) and the cytosol (**b**). Asterisks on heat maps indicate the subunits shown in the line plots.

the adaptation program; ATP synthase translation recovers over time, COB translation is not fully induced until 1 h and COX1 translation is not fully induced until 3 h. Thus, in contrast to the transcription of mitochondrial-encoded genes, mitochondrial translation is dynamically and differentially regulated.

Translation is coordinated across compartments

To determine whether cytosolic translation is coordinated with the rapid shift in mitochondrial translation efficiencies, we determined the relative synthesis and translation efficiencies for all cytosolic transcripts across our experimental conditions using cytoribosome profiling

(Supplementary Tables 4 and 5; representative library characteristics shown in Extended Data Fig. 4a, b). Abrupt transfer from a fermentable to a non-fermentable carbon source results in a transient reduction in cytosolic translation¹⁷ (Extended Data Fig. 7), but select transcripts escape this reduction and are preferentially translated to produce proteins that are required for cell survival under the new conditions²³. As expected, synthesis of all OXPHOS subunits increases as cells adapt to respiratory growth (Extended Data Fig. 5b); this increase is likely to be driven by the large-scale increase in RNA transcript levels that occurs immediately after carbon source shift. Remarkably, after normalizing for these RNA changes, the pattern of translational regulation of nuclear-encoded OXPHOS subunits is the same as for their mitochondrial-encoded counterparts (Fig. 3b). Within 15 min of nutrient shift, actively translating cytoribosomes markedly redistribute from ATP synthase mRNAs to complex III and complex IV mRNAs. Despite the double membrane separating cytoribosomes and mitoribosomes and a lack of any shared components, translation regulation of OXPHOS subunits is synchronized across cellular compartments.

Translation systems communicate unidirectionally

It is unclear whether communication between the two pools of ribosomes ensures the synchronized regulation of OXPHOS mRNAs, or whether each reacts independently to environmental signals. When mitochondria are isolated from cytosolic factors by treatment with an uncoupler (carbonyl cyanide *m*-chlorophenyl hydrazone (CCCP), which blocks mitochondrial import (Extended Data Fig. 8a)), mitochondrial translation is inhibited across all transcripts, independent of cytosolic translation and carbon source (Extended Data Fig. 8b–e and data not shown). Consistently, *in organello* mitochondrial translation is repressed unless the purified mitochondria are charged with a mixture of amino acids and nucleotides²⁴.

To determine directly whether communication occurs between the two translation systems, we specifically inhibited cytosolic translation with cycloheximide (CHX)²⁵ and observed the effect on mitochondrial translation (Fig. 4a–c and Extended Data Fig. 8b). CHX treatment does not affect mitochondrial mRNA levels (Extended Data Fig. 8c), and cells remain viable through the treatment course (Extended Data Fig. 8d). Upon complete inhibition of cytosolic translation, translation of

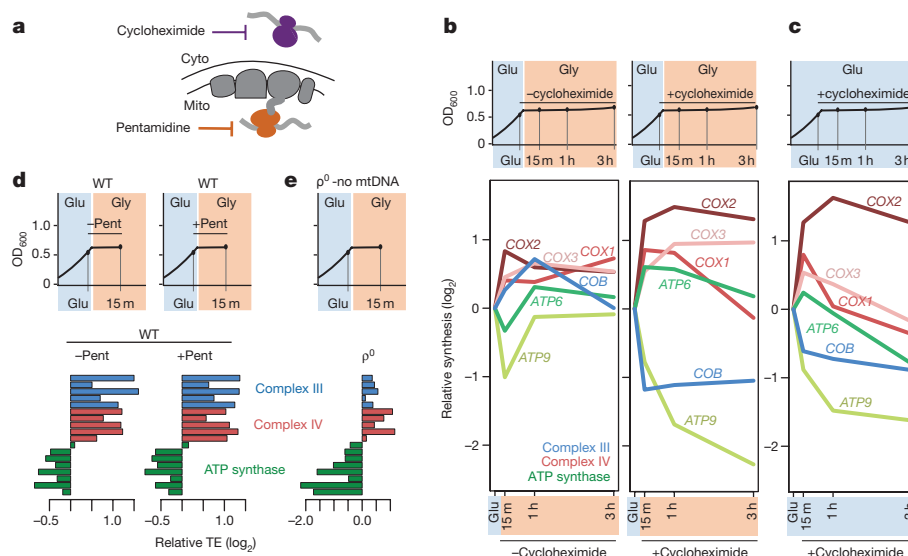


Figure 4 | Communication between translation systems is unidirectional. **a,** Schematic depicting action of drugs. **b, c,** Mitochondrial translation response, measured by northern blotting for footprints (Extended Data Fig. 8b), to cytosolic translation inhibition by CHX with (b) or without (c) carbon source shift. Note relative synthesis is a good proxy for translation efficiency (compare –CHX to Fig. 3a) because levels of mitochondrial mRNAs do not change substantially relative to

each other during this time period (Fig. 1c). Values in **b** are averages of two experiments. See Source Data for range values. **d, e,** Fold changes in translation efficiency of nuclear-encoded OXPHOS subunits measured by cytoribosome profiling (**d**) without (–Pent) or with (+Pent) inhibition of mitochondrial translation, and in ρ^0 cells (**e**). The subset of OXPHOS subunits shown is the same as that shown in line plots in Fig. 3b.

mitochondrial messages is differentially affected: some messages gain translational capacity and others lose it (Fig. 4b). Further, there is a decrease in the dynamics at later time points, with only minor changes occurring after the initial response; this suggests that ongoing cytosolic translation is important for changes in mitochondrial translation over the course of adaptation. Aside from *COB* and *ATP6*, CHX treatment does not substantially alter the rapid (15 min) mitochondrial translational response to a nutrient shift. Surprisingly, CHX treatment influences mitochondrial translation in a similar way even when no change in carbon source occurs (Fig. 4c). Thus, the early change in mitochondrial translation does not occur directly in response to environmental inputs, but rather is a reaction to the transient inhibition of cytosolic translation.

The ability of CHX to prevent translation of *COB* indicates that *COB* translation requires one or more newly synthesized cytosolic products that escape the global stress-induced inhibition of translation. Indeed, the synthesis of two *COB* translational activators, Cbp6 and Cbs2, is increased by nearly fourfold and sixfold, respectively, upon nutrient shift (Extended Data Fig. 8f). Consistent with the gain of *ATP6* translational capacity after treatment with CHX, the *ATP8/ATP6* transcript is the only one that has been found to have a translational repressor²⁶. These results suggest that cytosolic translational control of translational activators contributes to the orchestration of mitochondrial OXPHOS protein synthesis.

Having established that the synchronization of translational programs is actively controlled by cytosolic translation, we next investigated whether communication between the cytosol and mitochondria is unidirectional or bidirectional. We focused on the early response (15 min after nutrient shift), when translational changes are maximal (Fig. 3b) and any secondary responses should be minimized. When mitochondrial translation was specifically inhibited using pentamidine^{27,28} (Fig. 4a and Extended Data Fig. 9a, left panel), ribosome profiling revealed that cytoribosomes respond to the nutrient shift independently of mitochondrial translation (Fig. 4d and Extended Data Fig. 9b). To test the possibility that the cytosolic translation response on OXPHOS mRNAs results from feedback from alterations in membrane potential or the OXPHOS complexes themselves rather than mitochondrial translation, we created a ρ^0 yeast strain (Extended Data Fig. 10a, b). In this strain, which completely lacks mitochondrial DNA, mitochondrial translation, and functional OXPHOS complexes (Extended Data Fig. 9a, right panel), the nutrient shift-induced cytosolic translation response on OXPHOS mRNAs is maintained (Fig. 4e and Extended Data Fig. 9c). Thus, the synchronization of translation is facilitated through unidirectional communication from cytosolic to mitochondrial ribosomes.

Discussion

Translational reprogramming allows rapid changes in protein synthesis and conservation of cellular resources by focusing the expensive process of translation to where it is needed most^{29–31}. The preference for synthesis of complex III and IV subunits during adaptation to changes in carbon source may reflect the possibility that these subunits are present in fermenting cells at lower levels than ATP synthase, which functions in reverse in the absence of the electron transport chain to maintain the mitochondrial inner membrane potential². The core subunits of complexes III and IV (Cob and Cox1, respectively), are translationally upregulated at later time points compared to the early response at 15 min for the majority of subunits, probably owing to feedback mechanisms that couple their translation to assembly of their respective complexes^{32,33}. Thus, the synchronized translation could serve in part to maximize the efficiency of OXPHOS complex assembly and to limit nonproductive or harmful off-target interactions. Important goals for future studies include unravelling the roles of translation activators and other factors in mediating this synchronized response, and analysing metazoan systems in which some mitochondrial gene regulatory mechanisms have diverged from those found in yeast³. We expect that the

whole-cell genomic profiling approach described here will usher in an era of global gene expression analyses that will shed light on many aspects of mitochondrial biology in health and disease.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 29 October 2015; accepted 18 April 2016.

Published online 11 May 2016.

- Masters, B. S., Stohl, L. L. & Clayton, D. A. Yeast mitochondrial RNA polymerase is homologous to those encoded by bacteriophages T3 and T7. *Cell* **51**, 89–99 (1987).
- Faye, G. & Sor, F. Analysis of mitochondrial ribosomal proteins of *Saccharomyces cerevisiae* by two dimensional polyacrylamide gel electrophoresis. *Mol. Gen. Genet.* **155**, 27–34 (1977).
- Kehrein, K., Bonnefoy, N. & Ott, M. Mitochondrial protein synthesis: efficiency and accuracy. *Antioxid. Redox Signal.* **19**, 1928–1939 (2013).
- Bonitz, S. G. *et al.* Codon recognition rules in yeast mitochondria. *Proc. Natl Acad. Sci. USA* **77**, 3167–3170 (1980).
- Costanzo, M. C. & Fox, T. D. Control of mitochondrial gene expression in *Saccharomyces cerevisiae*. *Annu. Rev. Genet.* **24**, 91–113 (1990).
- Green-Willms, N. S., Butler, C. A., Dunstan, H. M. & Fox, T. D. Pet111p, an inner membrane-bound translational activator that limits expression of the *Saccharomyces cerevisiae* mitochondrial gene COX2. *J. Biol. Chem.* **276**, 6392–6397 (2001).
- Herrmann, J. M., Woellhaf, M. W. & Bonnefoy, N. Control of protein synthesis in yeast mitochondria: the concept of translational activators. *Biochim. Biophys. Acta* **1833**, 286–294 (2013).
- Müller, P. P. *et al.* A nuclear mutation that post-transcriptionally blocks accumulation of a yeast mitochondrial gene product can be suppressed by a mitochondrial gene rearrangement. *J. Mol. Biol.* **175**, 431–452 (1984).
- Fraenkel, D. G. *Yeast Intermediary Metabolism* (Cold Spring Harbor Press, 2011).
- Kuhn, K. M., DeRisi, J. L., Brown, P. O. & Sarnow, P. Global and specific translational regulation in the genomic response of *Saccharomyces cerevisiae* to a rapid transfer from a fermentable to a nonfermentable carbon source. *Mol. Cell. Biol.* **21**, 916–927 (2001).
- Amiott, E. A. & Jaehning, J. A. Mitochondrial transcription is regulated via an ATP “sensing” mechanism that couples RNA abundance to respiration. *Mol. Cell* **22**, 329–338 (2006).
- Mueller, D. M. & Getz, G. S. Steady state analysis of mitochondrial RNA after growth of yeast *Saccharomyces cerevisiae* under catabolite repression and derepression. *J. Biol. Chem.* **261**, 11816–11822 (1986).
- DeRisi, J. L., Iyer, V. R. & Brown, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
- Roberts, G. G. & Hudson, A. P. Transcriptome profiling of *Saccharomyces cerevisiae* during a transition from fermentative to glycerol-based respiratory growth reveals extensive metabolic and structural remodeling. *Mol. Genet. Genomics* **276**, 170–186 (2006).
- Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
- Fox, T. D. *et al.* Analysis and manipulation of yeast mitochondrial genes. *Methods Enzymol.* **194**, 149–165 (1991).
- Ashe, M. P., De Long, S. K. & Sachs, A. B. Glucose depletion rapidly inhibits translation initiation in yeast. *Mol. Biol. Cell* **11**, 833–848 (2000).
- Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).
- Rooijers, K., Loayza-Puch, F., Nijtmans, L. G. & Agami, R. Ribosome profiling reveals features of normal and disease-associated mitochondrial translation. *Nature Commun.* **4**, 2886 (2013).
- Vignais, P. V., Stevens, B. J., Huet, J. & Andre, J. Mitochondria from *Candida utilis*. Morphological, physical, and chemical characterization of the monomer form and of its subunits. *J. Cell Biol.* **54**, 468–492 (1972).
- Wolin, S. L. & Walter, P. Ribosome pausing and stacking during translation of a eukaryotic mRNA. *EMBO J.* **7**, 3559–3569 (1988).
- Poutre, C. G. & Fox, T. D. PET111, a *Saccharomyces cerevisiae* nuclear gene required for translation of the mitochondrial mRNA encoding cytochrome c oxidase subunit II. *Genetics* **115**, 637–647 (1987).
- Simpson, C. E. & Ashe, M. P. Adaptation to stress in yeast: to translate or not? *Biochem. Soc. Trans.* **40**, 794–799 (2012).
- Poyton, R. O., Bellus, G., McKee, E. E., Sevarino, K. A. & Goehring, B. *In organello* mitochondrial protein and RNA synthesis systems from *Saccharomyces cerevisiae*. *Methods Enzymol.* **264**, 36–42 (1996).
- Lamb, A. J., Clark-Walker, G. D. & Linnane, A. W. The biogenesis of mitochondria. 4. The differentiation of mitochondrial and cytoplasmic protein synthesizing systems *in vitro* by antibiotics. *Biochim. Biophys. Acta* **161**, 415–427 (1968).
- Rak, M. *et al.* Regulation of mitochondrial translation of the *ATP8/ATP6* mRNA by Smt1p. *Mol. Biol. Cell* (2016).
- Sun, T. & Zhang, Y. Pentamidine binds to tRNA through non-specific hydrophobic interactions and inhibits aminoacylation and translation. *Nucleic Acids Res.* **36**, 1654–1664 (2008).

28. Zhang, Y., Bell, A., Perlman, P. S. & Leibowitz, M. J. Pentamidine inhibits mitochondrial intron splicing and translation in *Saccharomyces cerevisiae*. *RNA* **6**, 937–951 (2000).
29. Liu, B. & Qian, S. B. Translational reprogramming in cellular stress response. *Wiley Interdiscip. Rev. RNA* **5**, 301–315 (2014).
30. Pavlov, M. Y. & Ehrenberg, M. Optimal control of gene expression for fast proteome adaptation to environmental change. *Proc. Natl Acad. Sci. USA* **110**, 20527–20532 (2013).
31. Sonenberg, N. & Hinnebusch, A. G. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* **136**, 731–745 (2009).
32. Gruschke, S. *et al.* The Cbp3-Cbp6 complex coordinates cytochrome *b* synthesis with *bc*₁ complex assembly in yeast mitochondria. *J. Cell Biol.* **199**, 137–150 (2012).
33. Perez-Martinez, X., Butler, C. A., Shingu-Vazquez, M. & Fox, T. D. Dual functions of Mss51 couple synthesis of Cox1 to assembly of cytochrome c oxidase in *Saccharomyces cerevisiae* mitochondria. *Mol. Biol. Cell* **20**, 4371–4380 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank F. Winston, T. Fox, G. Brar and members of the Churchman lab for advice and discussions; members of the O'Shea, Novina, and Springer labs for use of equipment and advice; and M. Hickman and

D. Botstein for sharing the *HAP1*⁺ strain. Research supported by a Damon Runyon Cancer Research Foundation Frey Award (to L.S.C.), a Burroughs Wellcome Fund Career Award at the Scientific Interface (to L.S.C.), an Ellison Medical Foundation New Scholar in Aging Award (to L.S.C.), the National Institutes of Health F32 (to M.T.C.), and a Boehringer Ingelheim Fonds PhD Fellowship (to G.S.).

Author Contributions M.T.C. and L.S.C. designed the research and wrote the manuscript. M.T.C. conducted the experiments with help from I.C.S., who performed mitoribosome profiling with drug treatments, and G.S., who created and performed mitoribosome profiling on the *pet111Δ* strain. M.T.C. analysed the data. All authors discussed the results and commented on the manuscript.

Author Information All data are deposited in Gene Expression Omnibus (accession number GSE74454). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.S.C. (churchman@genetics.med.harvard.edu).

Reviewer Information *Nature* thanks P. Van Damme and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments or outcome assessment.

Strain construction, growth conditions, and petite frequency analysis. To ensure a robust and physiological response, we modified the S288c background typically used for genomic studies to make it amenable to mitochondrial studies by correcting the *HAPI* mutation that leads to lower mitochondrial biomass production^{34,35} and creating a point change to restore a conserved amino acid that is important for the fidelity of the mitochondrial DNA (mtDNA) polymerase^{36,37}. These modifications markedly reduced the frequency of mtDNA loss (Extended Data Fig. 3b). S288c derivative DBY12045 (*MATa HAPI⁺ GAL2⁺ ura3Δ0 MIP1[S]*) was a gift from M. Hickman and D. Botstein. It was modified by generating a single point change using a loopout strategy to create *MIP1[S]^{A661T}*, repairing a strictly conserved threonine that is mutated in S288c and is responsible for the increased rate of mitochondrial mutations in this strain³⁶. Epitope-tagged proteins were expressed from their endogenous loci and generated using a 'scarless' loopout strategy^{38,39}. Deletion of *PET111* was performed using the *delitto perfetto* approach⁴⁰. For ρ^0 strain generation, mtDNA loss was induced by overnight growth in 10 $\mu\text{g ml}^{-1}$ ethidium bromide at 25°C.

Petite frequency was assayed essentially as described⁴¹. Briefly, fresh colonies from yeast extract peptone dextrose (YPD; 2% glucose) plates were resuspended and plated at low density on YPDG (0.1% glucose, 3% glycerol). 'Petite' and 'grande' colonies were counted after 5 days of growth at 30°C.

Yeast strains were grown in YP (1% yeast extract, 2% peptone) medium, pH 5.0, supplemented with 2% glucose (Glu) or 3% glycerol (Gly) as indicated. The *pet111Δ* strain is respiratory-deficient and was therefore grown in YPGal (2% galactose) instead of YPGly. Controlling the pH of the medium is essential for consistent growth on glycerol. Overnight liquid YPD cultures were grown to saturation and used to inoculate fresh medium to $\text{OD}_{600} \leq 0.06$. Cultures were grown at 30°C until OD_{600} reached 0.6–0.8. For rapid media transfer, cultures were harvested by filtration, rinsed once in YP (containing drug as indicated), and scraped off the filter into fresh medium. All media and flasks were pre-warmed. Where indicated, CHX (Sigma) was used in cultures at a final concentration of 100 $\mu\text{g ml}^{-1}$, pentamidine (Sigma) at 10 μM , and CCCP (Sigma) at 40 μM .

FACS analysis. Yeast cultures were grown until they reached $\text{OD}_{600} = 0.5$ in YP (1% yeast extract, 2% peptone), pH 5.0 supplemented with 2% glucose. Where indicated, cells were treated with 40 μM CCCP for 2 or 5 min. After treatment, the drug was washed off with 1× PBS and cells were diluted to 10⁶ cells per ml in 1× PBS. Where indicated, diluted cells were treated with tetramethylrhodamine (TMRM) at 1 μM for 30 min. Cells were washed twice, resuspended in 1× PBS to a final concentration of 10⁶ ml⁻¹ and loaded into 96-well culture plates (CellTreat). FACS analysis was performed using a Stratified 1000 instrument. Emission wavelengths were recorded at 586 nm. Histograms were generated using FlowJo software.

General nucleic acid and protein methods. For footprint detection by northern blotting, RNA was isolated from purified mitoribosomes (Flag eluate) and loaded on a 15% polyacrylamide TBE-urea gel. After transfer to nylon (Hybond N+), blots were probed at room temperature with internally labelled random hexamer-primed DNA fragments synthesized from 200–500-bp PCR-generated templates. For mRNA detection, RNA was separated on 1.2% formaldehyde agarose and blots were probed at 42°C. Quantitative PCR (qPCR) was performed using a CFX BioRad Connection qPCR thermocycler with EvaGreen (BioRad) fluorescent dye.

Proteins were resolved before silver staining or western blotting on NuPAGE Novex Bis-Tris gels (Thermo Fisher Scientific). Staining was done with the SilverQuest Silver Staining Kit (Invitrogen) according to the manufacturer's instructions. Antibodies against OXPHOS proteins were purchased from Santa Cruz Biotechnology (anti-Cob, sc-11436) and Abcam Mitosciences (anti-Cox1, ab110270; anti-Cox2, ab110271; anti-Cox4, ab110272). Fluorescently labelled secondary antibodies (IRDye, LI-COR) were detected using the LI-COR Odyssey.

Metabolic labelling was performed essentially as described⁴². Briefly, cultures were grown in YPGal (2% galactose) to log phase, washed and resuspended in potassium phosphate buffer at 30°C, shaking, for 2 h. Pentamidine was added to 10 μM where indicated and incubation continued for 15 min. To assay mitochondrial translation, CHX was added to 500 $\mu\text{g ml}^{-1}$ for 3 min before addition of ³⁵S-labelled methionine and cysteine mix (Perkin Elmer). Labelling was continued for 20 min at 30°C. Proteins were TCA precipitated, resolved on 17.5% Tris-glycine polyacrylamide, pH 8.3, and transferred to nitrocellulose. For visualizing cytosolic translation, CHX was omitted and protein samples were diluted 1:15 before loading the gel.

Mitoribosome profiling. Cell culture (400 OD_{600} equivalents) was rapidly harvested by filtration onto 0.45- μm nitrocellulose (Whatman). Cell pellets were flash frozen in liquid nitrogen and combined with 4 ml of frozen lysis buffer (10 mM Tris, pH 8.0, 50 mM NH_4Cl , 10 mM MgCl_2 , 0.5% lauryl maltoside, 0.25 mM DTT,

1.5× protease inhibitor cocktail (Complete, EDTA-free, Roche)). In optimizing buffer conditions to maintain mitoribosome subunit association, we found the ratio of monovalent to divalent cations to be of vital importance²⁰. The frozen cell mixture was pulverized in 50-ml canisters prechilled in liquid nitrogen for six cycles of 3 min each at 15 Hz, on a Retsch MM301 mixer mill. Upon thawing, fresh lysis buffer was added to bring the lysate concentration to 25 OD_{600} equivalents per ml. Lysate was digested for 30 min at 25°C with 500 U ml^{-1} of recombinant RNase I (RNase I₆, NEB). The reaction was stopped with 100 U ml^{-1} SUPERase-In (Thermo Fisher Scientific) and clarified by centrifugation at 4°C at 20,000g for 15 min. The supernatant was reserved for immunoprecipitation.

Anti-Flag M2 affinity gel (Sigma) was washed 3× with wash buffer (10 mM Tris pH 8.0, 50 mM NH_4Cl , 10 mM MgCl_2 , 0.1% Triton X-100), and added to clarified lysate at 12 μl 50% gel slurry per ml. The mixture was rotated end-over-end at 4°C for 3 h. The affinity gel was washed 3× for 10 min in 10 ml of wash buffer at room temperature, then Flag-tagged protein was eluted by incubation with 200 $\mu\text{g ml}^{-1}$ 3× Flag peptide (Sigma) in 6× volumes affinity gel slurry for 40 min at room temperature. RNA was isolated using phenol/chloroform extraction.

Mitoribosome-protected fragments were isolated by excision of ~36–42-nucleotide fragments from 12% or 15% TBE-urea polyacrylamide gels. Sequencing library preparation was performed through a circular intermediate as described⁴³, omitting the rRNA depletion step. Sequencing was performed on an Illumina MiSeq system using Reagent Kit v2 or v3. All experiments were performed in biological duplicate and means reported.

Cytoribosome profiling. Cytoribosome profiling for data presented in Fig. 3b and Extended Data Fig. 8f was performed using sucrose density gradients as described⁴³, except that CHX was omitted from cultures, and frozen cells were pulverized with lysis buffer containing CHX. Cytoribosome profiling for data presented in Fig. 4d, e and Extended Data Fig. 9b, c was performed using 1.0 M sucrose cushions in place of gradients. Sequencing libraries were prepared identically to those for mitoribosome profiling (above). Sequencing was performed on an Illumina NextSeq 500 system. Each experiment presented was performed once as cytoribosome profiling is highly reproducible^{44,45}. Additionally, trends are reproducible between experiments performed with density gradients and cushions.

mRNA sequencing. Total RNA was isolated before RNase I digestion from a portion of the thawed lysates described above. ERCC RNA Spike-In Mix (Thermo Fisher Scientific), a mixture of 92 *in vitro* synthesized transcripts, was added in equal volumes across samples that were prepared from equal cell numbers. 50 μg of the total RNA with spike-in mix was digested with 3 U RQ1 RNase-free DNase (Promega) for 30 min at 37°C. 5 μg of purified RNA was then subjected to rRNA depletion using the Ribo-Zero Magnetic Gold Kit for yeast (Epicentre) according to the manufacturer's instructions.

Sequencing libraries were prepared as above following fragmentation by alkaline hydrolysis in 5 mM Na_2CO_3 , 45 mM NaHCO_3 , 1 mM EDTA, pH 9.3 for 25 min at 95°C and gel isolation of 30–70-nucleotide fragments unless otherwise noted (Extended Data Fig. 4b). Sequencing was performed on an Illumina NextSeq 500 system. Each experiment was performed once.

Data analysis. Raw sequences were processed by first removing 3' adaptor sequence using Cutadapt⁴⁶ and removing the first nucleotide from the 5' end of all reads (except where otherwise noted) because we observed, as has been previously reported⁴³, that this nucleotide frequently represents untemplated addition by reverse transcriptase Superscript III (Thermo Fisher). Next, reads mapping to non-coding RNAs were removed by aligning using Bowtie1 (ref. 47) to a collection of RNA genes downloaded from the *Saccharomyces* Genome Database. Notably, we allowed a 3-nt 3' mismatch when mapping to tRNAs to account for non-templated CCA addition on mature tRNAs. Remaining reads were then aligned allowing two mismatches to the *S. cerevisiae* genome assembly R64 (UCSC: sacCer3) using Tophat2 (ref. 48). We also aligned separately to the nuclear and mitochondrial genomes to determine the proportions of each library mapping to each.

To determine the A-site position in mitoribosome footprints we observed the first reads mapping to each ORF, which overlap the start codons. Consistently for 36–40-nucleotide reads, the 3' ends of these reads were 19 nucleotides downstream of the start codon (mitoribosome P site). The 5' ends were more heterogeneous, depending on length of the read. Thus we assigned A sites for each read as 16 nucleotides from the 3' end.

RNA-seq normalization to spike-ins was performed by dividing the raw read count at each position by the number of spike-in reads in the library (RPS, reads per spike-in). The number of spike-in reads is a proxy for cell number (see above). RNA-seq reads were also normalized to total nuclear or mitochondrial mRNA mappers, as were cytoribosome profiling and mitoribosome profiling reads, respectively (RPM).

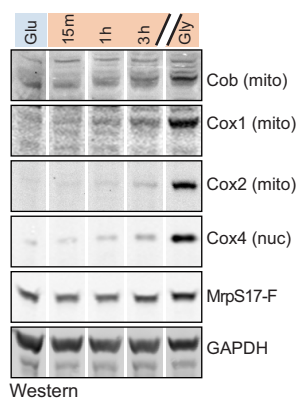
To determine expression values for each gene, RPS or RPM values were summed across ORFs and normalized to ORF length (RPKS or RPKM, respectively; normalized reads per kb). For footprint reads, the first and last five codons were

excluded to remove effects of translation initiation and termination⁴⁹. Translation efficiency was calculated by dividing cytoribosome footprint RPKM values by nuclear-mapping RNA-seq RPKM values and mitoribosome footprint RPKM values by mito-mapping RNA-seq RPKM values.

Scripts for A-site assignment, normalization, translation efficiency calculation, and other text file manipulations were written for Python 2.7.5. Plots and genome browser visualization were generated using R version 3.2.2 and Bioconductor. Heat maps were generated with Matrix2png (<http://www.chibi.ubc.ca/matrix2png/>).

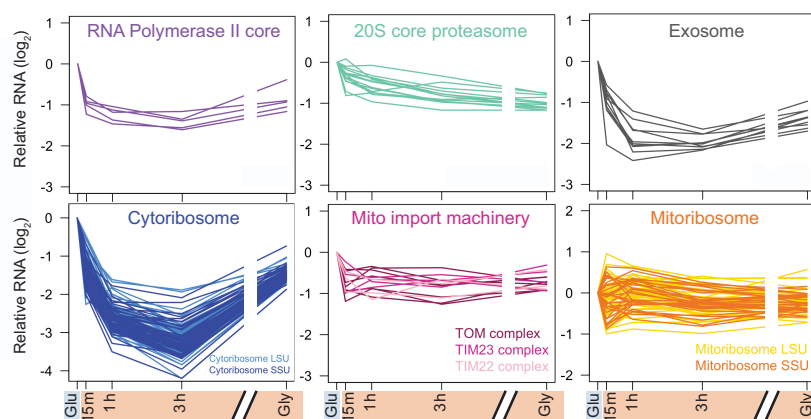
For analysis of mitoribosome footprints by northern blotting, phosphorimager signals were quantified using ImageJ (<http://imagej.nih.gov/>) and normalized to the amount of mitoribosome recovered as measured by Mrps17-Flag in elution. Comparisons were made only between samples on the same membrane probed at the same time with the same probe.

34. Ehrenreich, I. M. *et al.* Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature* **464**, 1039–1042 (2010).
35. Gaisne, M., Becam, A. M., Verdier, J. & Herbert, C. J. A 'natural' mutation in *Saccharomyces cerevisiae* strains derived from S288c affects the complex regulatory gene *HAP1* (*CYP1*). *Curr. Genet.* **36**, 195–200 (1999).
36. Baruffini, E., Lodi, T., Dallabona, C. & Foury, F. A single nucleotide polymorphism in the DNA polymerase gamma gene of *Saccharomyces cerevisiae* laboratory strains is responsible for increased mitochondrial DNA mutability. *Genetics* **177**, 1227–1231 (2007).
37. Young, M. J. & Court, D. A. Effects of the S288c genetic background and common auxotrophic markers on mitochondrial DNA function in *Saccharomyces cerevisiae*. *Yeast* **25**, 903–912 (2008).
38. Alani, E., Cao, L. & Kleckner, N. A method for gene disruption that allows repeated use of URA3 selection in the construction of multiply disrupted yeast strains. *Genetics* **116**, 541–545 (1987).
39. Moqtaderi, Z. & Struhl, K. Expanding the repertoire of plasmids for PCR-mediated epitope tagging in yeast. *Yeast* **25**, 287–292 (2008).
40. Storici, F. & Resnick, M. A. The delitto perfetto approach to *in vivo* site-directed mutagenesis and chromosome rearrangements with synthetic oligonucleotides in yeast. *Methods Enzymol.* **409**, 329–345 (2006).
41. Dimitrov, L. N., Brem, R. B., Kruglyak, L. & Gottschling, D. E. Polymorphisms in multiple genes contribute to the spontaneous mitochondrial genome instability of *Saccharomyces cerevisiae* S288C strains. *Genetics* **183**, 365–383 (2009).
42. Gouget, K., Verde, F. & Barrientos, A. *In vivo* labeling and analysis of mitochondrial translation products in budding and in fission yeasts. *Methods Mol. Biol.* **457**, 113–124 (2008).
43. Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M. & Weissman, J. S. The ribosome profiling strategy for monitoring translation *in vivo* by deep sequencing of ribosome-protected mRNA fragments. *Nature Protocols* **7**, 1534–1550 (2012).
44. Brar, G. A. *et al.* High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* **335**, 552–557 (2012).
45. Ingolia, N. T. Genome-wide translational profiling by ribosome footprinting. *Methods Enzymol.* **470**, 119–142 (2010).
46. Martin, M. Cutadapt removes adaptor sequences from high-throughput sequencing reads. *EMBnet journal* **17**, 10–12 (2011).
47. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
48. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
49. Li, G. W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* **157**, 624–635 (2014).

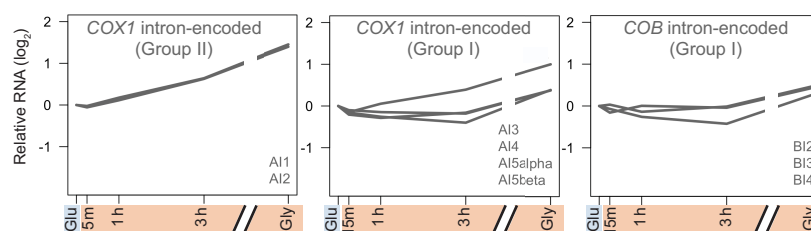


Extended Data Figure 1 | OXPHOS proteins are induced during mitochondrial biogenesis. Western blot analysis of mitochondrial (Cob, Cox1, Cox2) and nuclear (Cox4) OXPHOS proteins compared to Flag-tagged mitoribosome small subunit protein MrpS17. GAPDH was used as a loading control. For gel source data, see Supplementary Fig. 1.

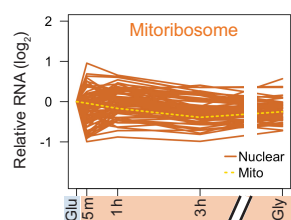
a



b

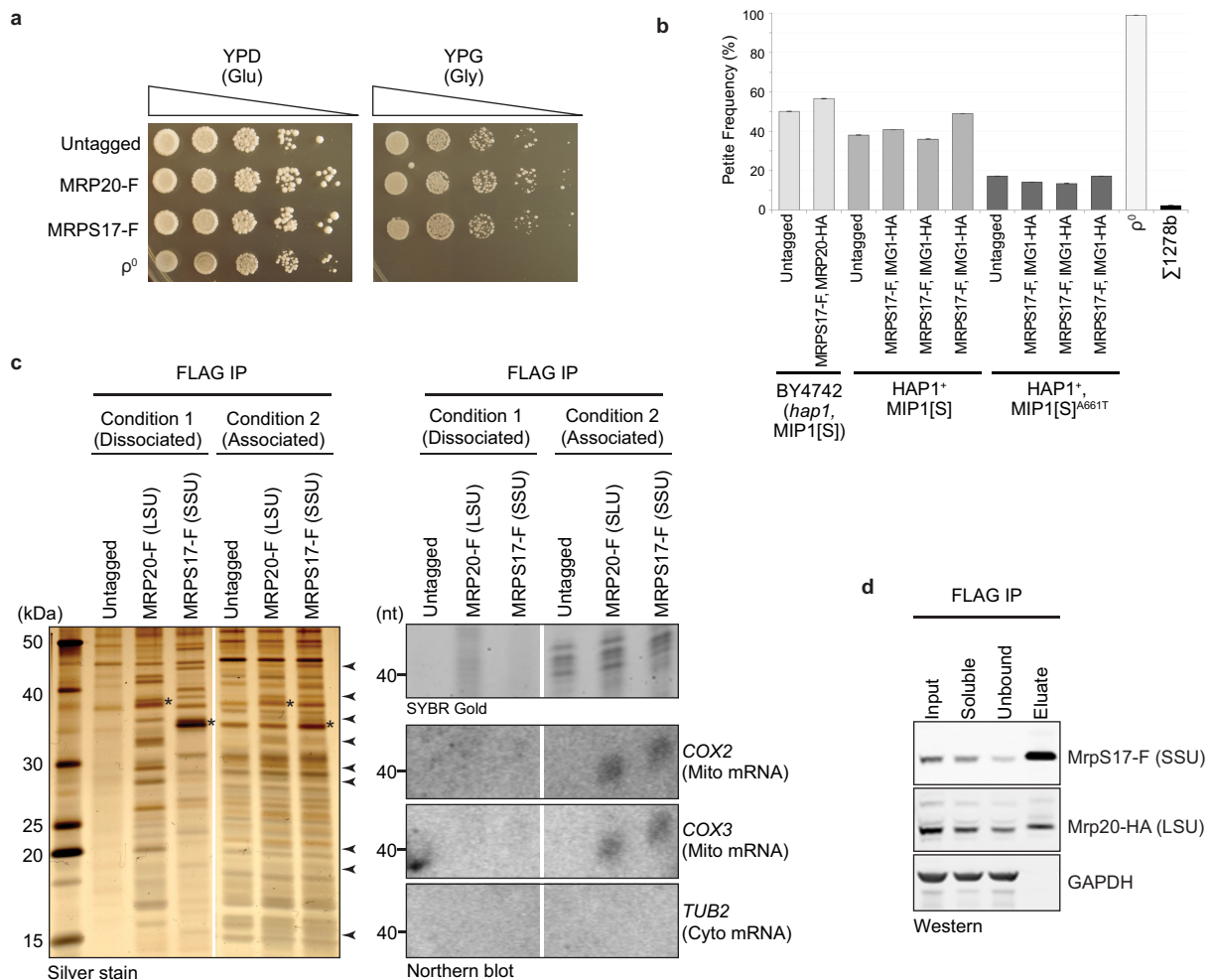


C



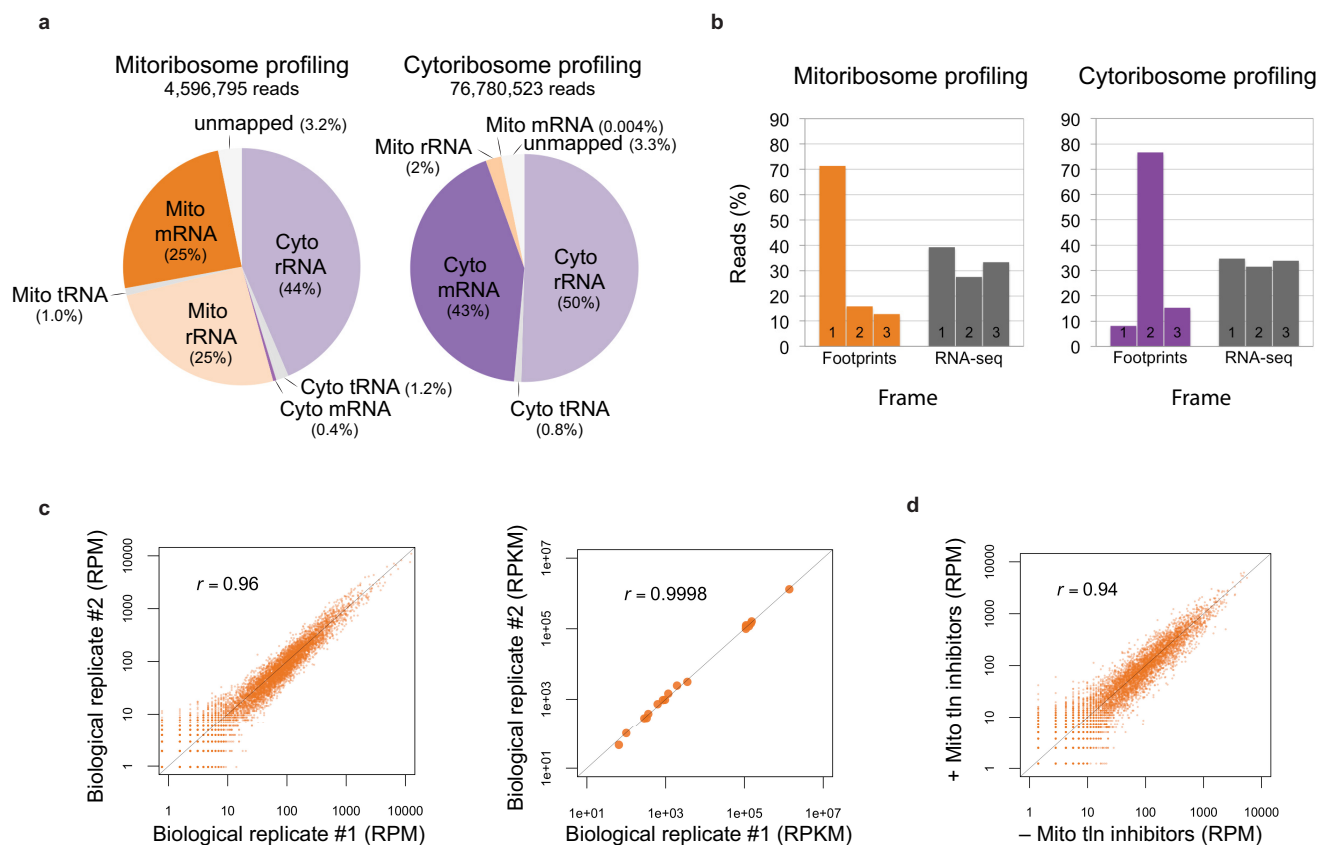
Extended Data Figure 2 | Dynamics of non-OXPHOS RNAs during mitochondrial biogenesis. **a–c**, RNA levels (reads per kb) normalized to spike-in controls and plotted as fold changes compared to levels in log phase glucose growth for all nuclear-encoded structural components of the complexes shown (**a**); intron-encoded maturases (**b**); and nuclear and

mitochondrial-encoded mitoribosome subunits (c). To calculate values for maturase transcripts, only reads not overlapping the main ORF (*COX1* or *COB*) were considered. Group II intron splicing intermediates stably accumulate and may not represent translation-competent transcripts.



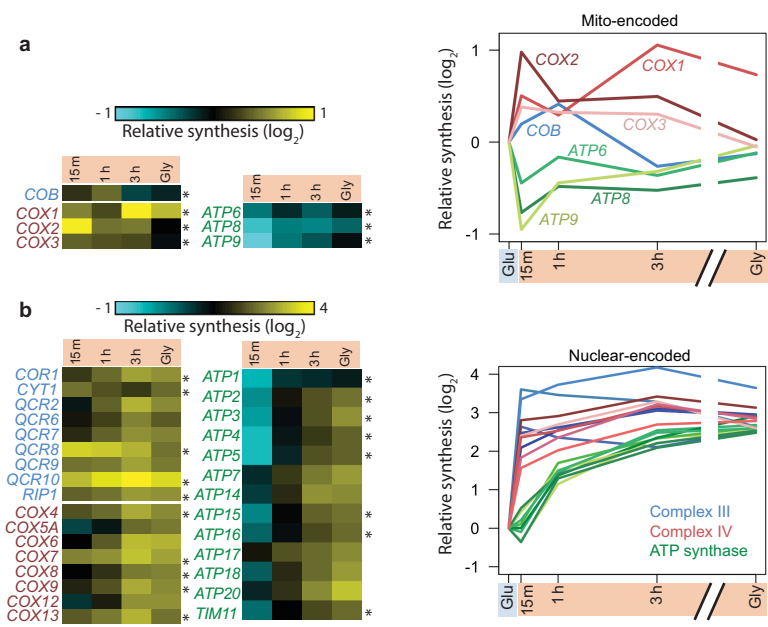
Extended Data Figure 3 | Optimization of affinity purification for intact mitoribosomes. **a**, Serial dilution spot tests verifying tagged mitoribosome subunits are functional as they support respiratory growth on glycerol (YPG). ρ^0 is a strain without mitochondrial DNA. **b**, Frequency of petite colonies in our corrected S288c strain (see Methods) after growth for 5 days on 0.1% glucose and 3% glycerol. BY4742 is S288c background with designer auxotrophies. $\Sigma 1278b$ is a strain with wild-type *HAP1*, and a high-fidelity allele of *MIPI*, *MIPI*[Σ], along with other differences compared to S288c. Error bars show variation due to counting, with 175–750 colonies counted for each sample. **c**, Lysis and immunoprecipitation buffer conditions affect mitoribosome subunit association and thus footprint retention. Left: silver staining after immunoprecipitation of the large subunit (LSU) with MRP20–Flag and of the small subunit

(SSU) with Mrps17–Flag in condition 1 (20 mM Tris pH 8.0, 200 mM KCl, 5 mM $MgCl_2$, 0.5% lauryl maltoside), and in condition 2 (10 mM Tris pH 8.0, 50 mM NH_4Cl , 10 mM $MgCl_2$, 0.5% lauryl maltoside). Arrowheads indicate bands that appear in both immunoprecipitations in condition 2 that can be assigned to the LSU or SSU by comparison to condition 1. Asterisks mark the expected mobility of the tagged proteins. Right: northern blotting of the co-purifying RNA in each condition. For gel source data, see Supplementary Fig. 1. **d**, Western blot showing fractions from immunoprecipitation using optimized buffer conditions. Flag immunoprecipitation targeting the mitoribosome SSU co-purifies a haemagglutinin (HA)-tagged LSU protein. For gel source data, see Supplementary Fig. 1.



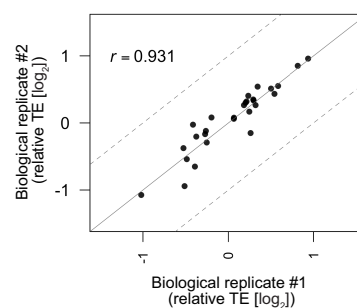
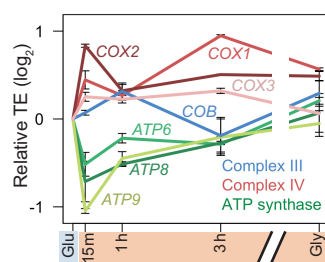
Extended Data Figure 4 | Mitoribosome profiling is robust, reproducible, and does not require translation inhibitors. **a**, Mapping statistics for representative mitoribosome and cytoribosome profiling libraries from log phase glycerol-grown cells. **b**, Fraction of reads mapping to each frame of mitochondrial ORFs (left) and nuclear ORFs (right) in mitoribosome profiling and cytoribosome profiling data, respectively. RNA-seq reads in the left panel were treated identically to footprint

reads, including size selection for library generation. **c**, Reproducibility between biological replicates. Each dot corresponds to the number of reads mapped to a particular position on mRNA (RPM, left), or summed number of reads mapped across each mRNA then normalized to length (RPKM, right). **d**, Reproducibility with and without translation inhibitors thiamphenicol ($50 \mu\text{g ml}^{-1}$) and GTP analogue GMPPNP (1 mM).



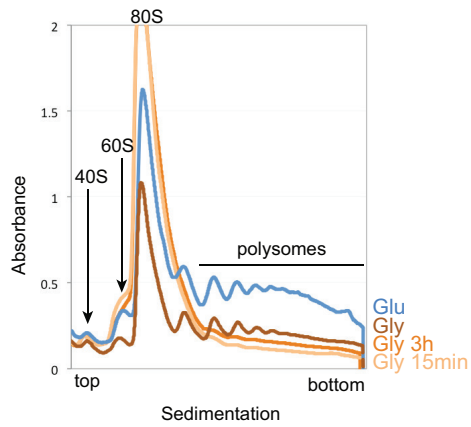
Extended Data Figure 5 | Mitochondrial and cytosolic protein synthesis on OXPHOS mRNAs is rapidly regulated. **a, b**, Fold changes in relative protein synthesis (footprint RPKM values) compared to log phase glucose

growth for the OXPHOS subunits synthesized in mitochondria (**a**; values are means of two experiments) and cytosol (**b**). Asterisks on heat maps indicate the subunits shown in the line plots.

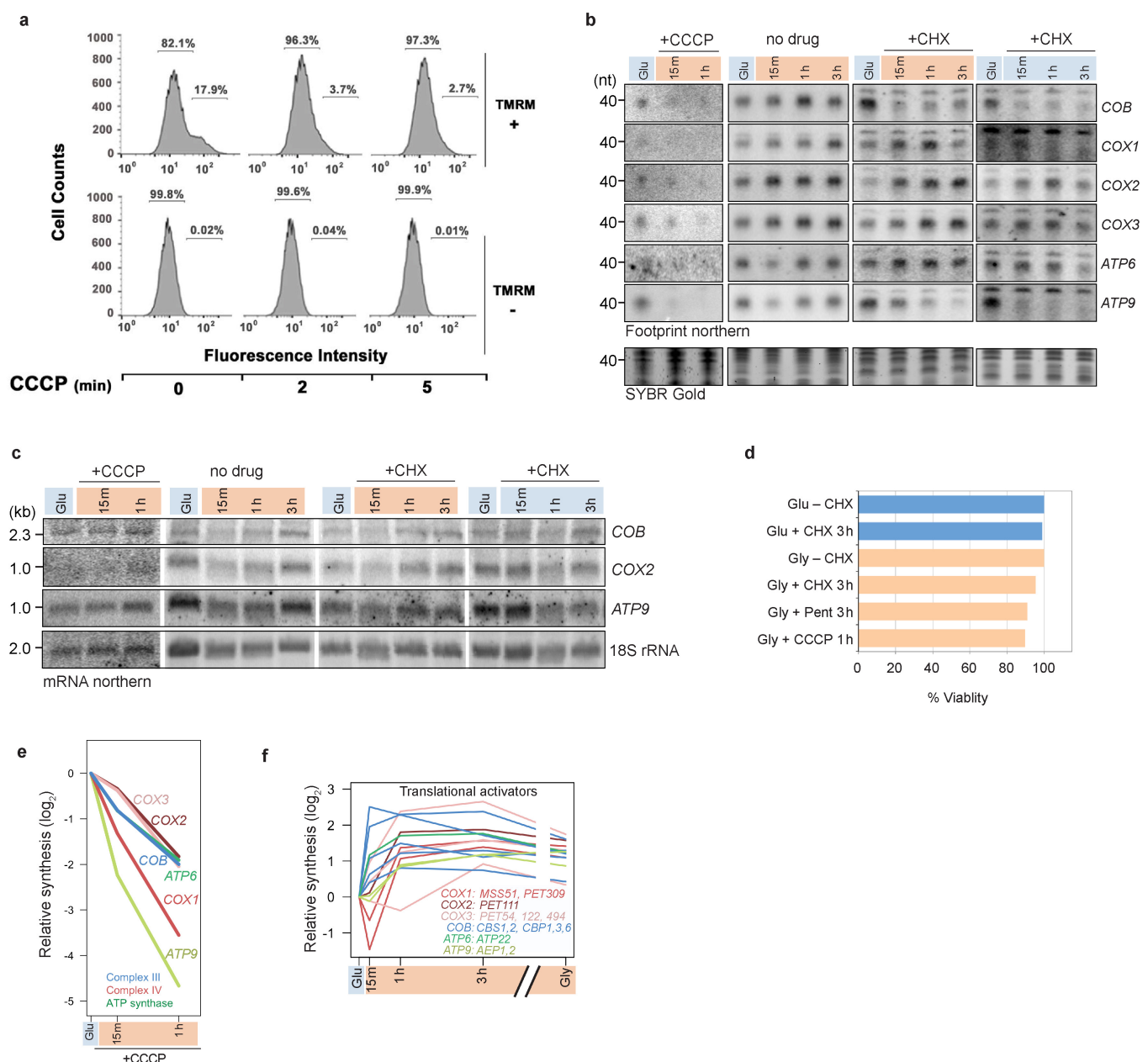


Extended Data Figure 6 | Mitochondosome translation efficiency fold changes are reproducible. Fold change data identical to those shown in Fig. 3a, but including range bars for two experiments performed from independent cultures on different days (left), and fold change translation

efficiency data plotted as a scatter with the Pearson correlation coefficient (right). Dotted lines mark twofold difference. RNA-seq data used in calculating translation efficiency are from a single experiment.

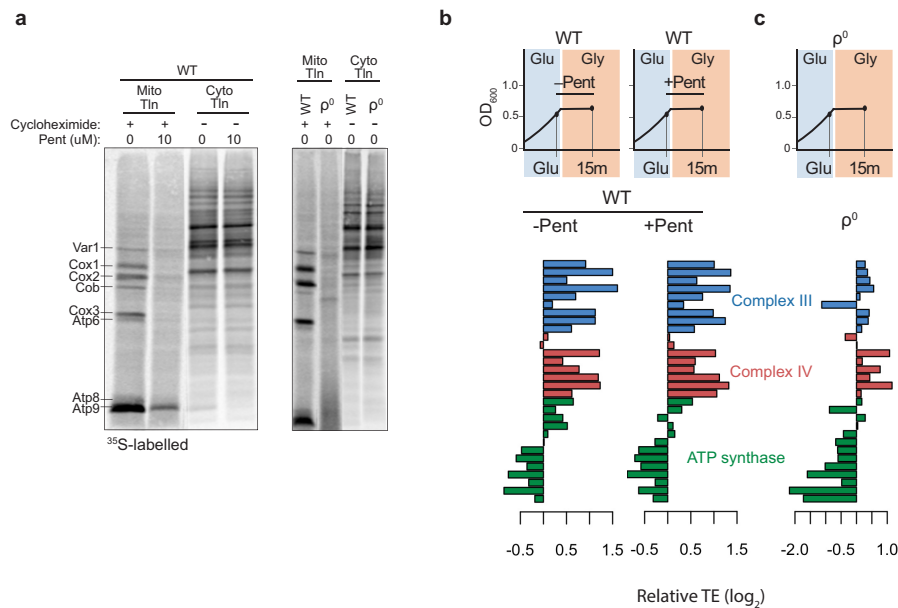


Extended Data Figure 7 | Global translation is transiently inhibited upon shift to glycerol. Polysome profiles from samples used for cytoribosome profiling, but without RNase I treatment. Gradients were loaded with lysate from equal cell numbers, allowing overall ribosome abundances to be compared between samples. Doubling time during log phase is ~ 1.2 h in glucose and ~ 3.7 h in glycerol.



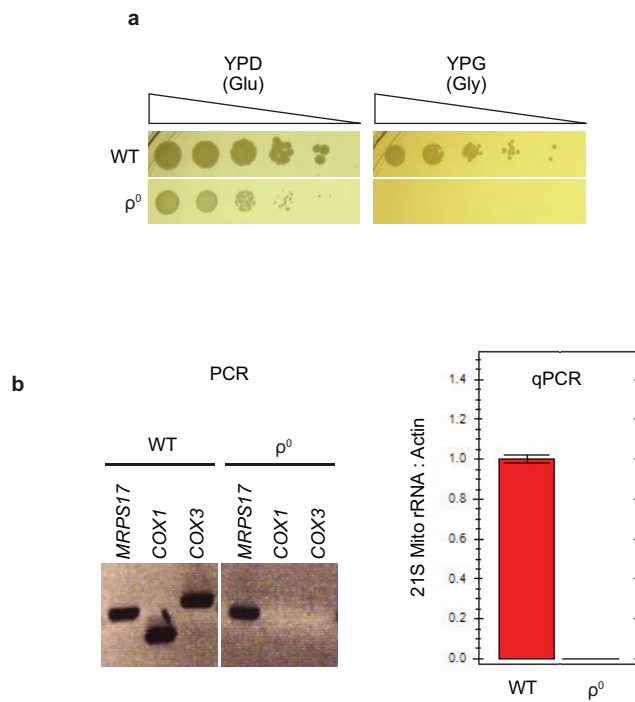
Extended Data Figure 8 | Cytosolic translation controls mitochondrial translation response. **a**, FACS analysis of yeast cultures treated with CCCP. Wild-type cultures were grown in glucose to mid-log phase and treated with 40 μ M CCCP for the indicated times. Mitochondrial membrane potential ($\Delta\Psi$) was assessed using 1 μ M tetramethylrhodamine (TMRM), which is taken up by only a fraction of the cell population (17.9% in this experiment). TMRM accumulates inside negatively charged mitochondria, producing increased fluorescence intensity (10^2). Loss of membrane potential dissipates probe, measured as loss of high-intensity fluorescence. **b**, Representative northern blots for data in Fig. 4b, c and panel **e**. For quantification, northern signals were normalized by relative mitoribosome recovery measured by Mrps17-Flag signal in western blots. For gel source data, see Supplementary Fig. 1.

c, Northern blotting of total RNA for the indicated transcripts. For gel source data, see Supplementary Fig. 1. **d**, Quantification of viability assay. Cells were grown in YPD (Glu) or YPG (Gly) with or without drug for the time indicated. Cells were washed out of drug and plated on YPD for calculation of colony-forming units. CCCP (40 μ M); CHX (100 μ g ml⁻¹); pentamidine (Pent; 10 μ M). **e**, Mitochondrial translation response to inhibition of mitochondrial import with CCCP, measured by northern blotting for footprints (see **b**). **f**, Fold change in synthesis measured by cytoribosome profiling of nuclear-encoded mitochondrial mRNA-specific translational activators (colour-coded by mRNA target). For each mitochondrial mRNA, the names of the known translational activators are listed.



Extended Data Figure 9 | Cytosolic OXPHOS translation response is independent of mitochondrial gene expression. a, Metabolic labelling to measure mitochondrial translation (Mito Tln), detectable only in the presence of CHX, and cytosolic translation (Cyto Tln). Samples generated in the absence of CHX were diluted 15-fold before loading the gel compared to samples generated with CHX. Mitochondrial translation

products are labelled. **b, c,** Full data set for experiment presented in Fig. 4d, e, showing fold changes in translation efficiency of all nuclear-encoded complex III, complex IV, and ATP synthase subunits measured by cytoribosome profiling without (–Pent) or with (+Pent) inhibition of mitochondrial translation (**b**) and in ρ^0 cells (**c**), which have neither mitochondrial translation nor functional OXPHOS complexes.



Extended Data Figure 10 | Verification of mtDNA loss in ρ^0 strain.

a, Spot tests verifying that the ρ^0 strain generated by overnight growth in ethidium bromide (see Methods) cannot respire (no growth on YPG).

b, PCR (left) and qPCR (right) verifying loss of mitochondrial-encoded genes *COX1*, *COX3*, and the 21S mitochondrial rRNA gene. *MRPS17* is nuclear-encoded. Bars show s.e.m. for technical triplicates.

Suppressing star formation in quiescent galaxies with supermassive black hole winds

Edmond Cheung¹, Kevin Bundy¹, Michele Cappellari², Sébastien Peirani^{1,3}, Wiphu Rujopakarn^{1,4}, Kyle Westfall⁵, Renbin Yan⁶, Matthew Bershad⁷, Jenny E. Greene⁸, Timothy M. Heckman⁹, Niv Drory¹⁰, David R. Law¹¹, Karen L. Masters⁵, Daniel Thomas⁵, David A. Wake^{7,12}, Anne-Marie Weijmans¹³, Kate Rubin¹⁴, Francesco Belfiore^{15,16}, Benedetta Vulcani¹, Yan-mei Chen¹⁷, Kai Zhang⁶, Joseph D. Gelfand^{18,19}, Dmitry Bizyaev^{20,21}, A. Roman-Lopes²² & Donald P. Schneider^{23,24}

Quiescent galaxies with little or no ongoing star formation dominate the population of galaxies with masses above 2×10^{10} times that of the Sun; the number of quiescent galaxies has increased by a factor of about 25 over the past ten billion years (refs 1–4). Once star formation has been shut down, perhaps during the quasar phase of rapid accretion onto a supermassive black hole^{5–7}, an unknown mechanism must remove or heat the gas that is subsequently accreted from either stellar mass loss⁸ or mergers and that would otherwise cool to form stars^{9,10}. Energy output from a black hole accreting at a low rate has been proposed^{11–13}, but observational evidence for this in the form of expanding hot gas shells is indirect and limited to radio galaxies at the centres of clusters^{14,15}, which are too rare to explain the vast majority of the quiescent population¹⁶. Here we report bisymmetric emission features co-aligned with strong ionized-gas velocity gradients from which we infer the presence of centrally driven winds in typical quiescent galaxies that host low-luminosity active nuclei. These galaxies are surprisingly common, accounting for as much as ten per cent of the quiescent population with masses around 2×10^{10} times that of the Sun. In a prototypical example, we calculate that the energy input from the galaxy's low-level active supermassive black hole is capable of driving the observed wind, which contains sufficient mechanical energy to heat ambient, cooler gas (also detected) and thereby suppress star formation.

Using optical imaging spectroscopy from the Sloan Digital Sky Survey-IV Mapping Nearby Galaxies at Apache Point Observatory¹⁷ (SDSS-IV MaNGA) programme, we define a new class of quiescent galaxies—selected to have red rest-frame colours, $\text{NUV} - r > 5$, where NUV and r are the magnitudes in the near-ultraviolet and r band, respectively—that is characterized by the presence of narrow bisymmetric patterns in equivalent width (EW) maps of strong emission lines, such as H α and [O III]. Our selection employs multiband imaging to exclude galaxies with dust lanes and other disk signatures. The observed enhanced emission features are oriented randomly with respect to the optical surface brightness morphology, but roughly align with strong, systematic velocity gradients as traced by the ionized gas emission lines. The gas velocity fields in these galaxies are decoupled

from their stellar motions. These galaxies are surprisingly common among the quiescent population, accounting for $\sim 10\%$ of quiescent galaxies with $\log(M_*/M_\odot) \approx 10.3$ (here M_* is the stellar mass and M_\odot is the solar mass).

To illuminate the salient features of this class, we focus on a prototypical example, informally named 'Akira' (Fig. 1). The SDSS imaging shows Akira to be an unremarkable spheroidal galaxy of moderate stellar mass ($\log(M_*/M_\odot) = 10.78$) that is interacting with a low-mass companion (informally named 'Tetsuo') at a projected separation of ~ 32 kpc ($67''$); they are not classified as members of a larger galaxy group¹⁸ and the properties of both galaxies are listed in Table 1. Spectral energy distribution (SED) fitting indicates that Akira is nearly dormant, with almost no detection of ongoing star formation¹⁹. Spatially resolved spectroscopy, however, reveals intriguing and complex patterns among spectral tracers of gas in Akira that point to a much more active internal state. With ionized-gas emission detected across the entire galaxy, the map of H α EW (which measures the line flux relative to the stellar continuum; Fig. 1c) reveals a prominent and somewhat twisted bisymmetric pattern with a position angle (PA) of $\sim 46^\circ$. The projected velocity gradient of ionized gas ranges from $v_{\text{ionized gas}} = -225 \text{ km s}^{-1}$ to $v_{\text{ionized gas}} = 200 \text{ km s}^{-1}$ along the kinematic major axis, which is at a PA of $\sim 26^\circ$ (Fig. 1h). We observe high ionized-gas velocity dispersions ($\sigma_{\text{ionized gas}}$) across the galaxy with interesting internal structure and maxima that reach $\sigma_{\text{ionized gas}} \approx 200 \text{ km s}^{-1}$ (Fig. 1i) and line widths of $W_{80} \approx 500 \text{ km s}^{-1}$ (see Methods) perpendicular to the major kinematic axis. Meanwhile, stellar motions reveal a minimal gradient ($\pm 30 \text{ km s}^{-1}$; Fig. 1f) that follows the PA of the galaxy's elliptical isophotes of $\sim 53^\circ$ (contours in Fig. 1c). We also detect a spatially offset enhancement in Na D absorption (Fig. 1d) that is coincident with excess dust in our derived extinction map (see Methods). Measurements of the Na D line centre trace a separate and distinct velocity gradient field across the offset absorption (Fig. 1e) that ranges from approximately $v_{\text{Na D}} = -80 \text{ km s}^{-1}$ to $v_{\text{Na D}} = 60 \text{ km s}^{-1}$.

These observations indicate the presence of multiple gas components with different temperatures and velocity structures. We interpret the ionized-gas velocity field as resulting from a centrally driven (volume-filling) wind with a wide opening angle. The projected flux

¹Kavli Institute for the Physics and Mathematics of the Universe (World Premier International Research Center Initiative), The University of Tokyo Institutes for Advanced Study, The University of Tokyo, Kashiwa, Chiba 277-8583, Japan. ²Sub-department of Astrophysics, Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK. ³Institut d'Astrophysique de Paris (UMR 7095, CNRS and UPMC), 98 bis Boulevard Arago, F-75014 Paris, France. ⁴Department of Physics, Faculty of Science, Chulalongkorn University, 254 Phayathai Road, Pathumwan, Bangkok 10330, Thailand. ⁵Institute for Cosmology and Gravitation, University of Portsmouth, Dennis Sciama Building, Burnaby Road, Portsmouth PO1 3FX, UK. ⁶Department of Physics and Astronomy, University of Kentucky, 505 Rose Street, Lexington, Kentucky 40506-0055, USA. ⁷Department of Astronomy, University of Wisconsin-Madison, 475 North Charter Street, Madison, Wisconsin 53706, USA. ⁸Department of Astrophysical Sciences, Princeton University, Princeton, New Jersey 08544, USA. ⁹Center for Astrophysical Sciences, Department of Physics and Astronomy, The Johns Hopkins University, Baltimore, Maryland 21218, USA. ¹⁰McDonald Observatory, Department of Astronomy, University of Texas at Austin, 1 University Station, Austin, Texas 78712-0259, USA. ¹¹Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, Maryland 21218, USA. ¹²Department of Physical Sciences, The Open University, Milton Keynes MK7 6AA, UK. ¹³School of Physics and Astronomy, University of St Andrews, North Haugh, St Andrews, Fife KY16 9SS, UK. ¹⁴Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA. ¹⁵Cavendish Laboratory, University of Cambridge, 19 J. J. Thomson Avenue, Cambridge CB3 0HE, UK. ¹⁶Kavli Institute for Cosmology, University of Cambridge, Cambridge CB3 0HE, UK. ¹⁷Department of Astronomy, Nanjing University, Nanjing 210093, China. ¹⁸New York University Abu Dhabi, PO Box 129188, Abu Dhabi, United Arab Emirates. ¹⁹Center for Cosmology and Particle Physics, New York University, Meyer Hall of Physics, 4 Washington Place, New York, New York 10003, USA. ²⁰Apache Point Observatory and New Mexico State University, PO Box 59, Sunspot, New Mexico 88349-0059, USA. ²¹Sternberg Astronomical Institute, Moscow State University, Moscow, Russia. ²²Departamento de Física y Astronomía, Facultad de Ciencias, Universidad de La Serena, Cisternas 1200, La Serena, Chile. ²³Department of Astronomy and Astrophysics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. ²⁴Institute for Gravitation and the Cosmos, The Pennsylvania State University, University Park, Pennsylvania 16802, USA.

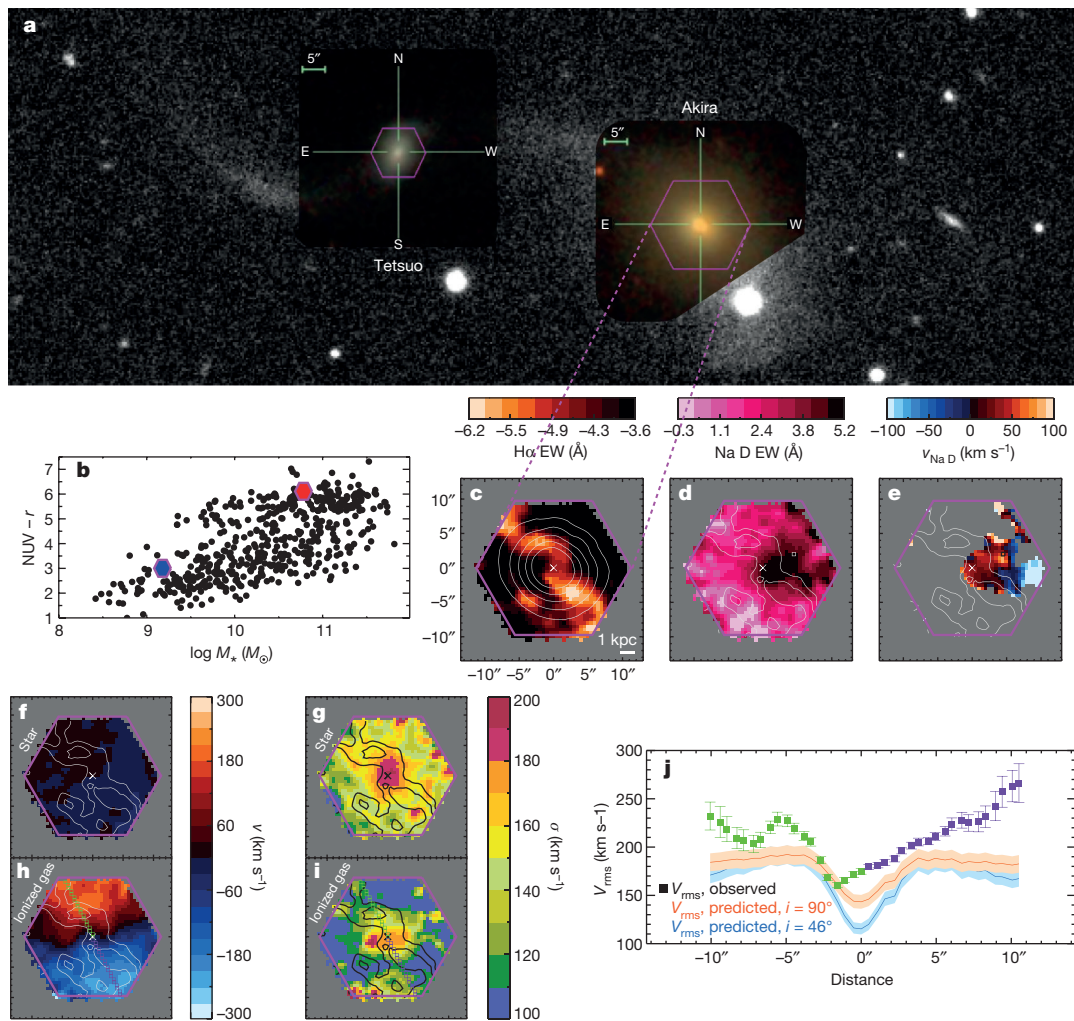


Figure 1 | Akira is the prototypical red geyser. **a**, The SDSS *gri* colour images of Akira (West) and Tetsuo (East) embedded in a larger SDSS *r* image, with the MaNGA footprint in pink. **b**, The rest-frame NUV – *r* versus $\log M_*$ diagram of the adopted MaNGA sample, with Akira and Tetsuo highlighted. **c**, The H α EW, with contours tracing the stellar continuum. **d**, The Na D EW. **e**, The Na D velocity. **f**, The stellar velocity. **g**, The stellar velocity dispersion. **h**, The ionized gas velocity. **i**, The

ionized gas velocity dispersion. The H α EW contours are overplotted on panels **d**–**i**. **j**, The observed V_{rms} from the highlighted spaxels (spectral pixels) exceeds the V_{rms} predicted from the gravitational potential, ruling out disk-like rotation. Error bars on the observed V_{rms} represent the 1σ measurement errors, while shaded regions around the predicted V_{rms} represent a conservative estimate of the systematic uncertainties.

distribution of this ionized component largely follows the stellar surface brightness, suggesting that its primary ionization source is the local radiation field from evolved stars^{20–22}. The bisymmetric EW features represent enhanced emission due to shocks or over-densities along the wind's central axis. A distinct and cooler gas component is indicated by the Na D absorption. Because it is spatially confined with its own velocity structure, this cooler foreground material is likely to be within 1–2 effective radius (R_e) of Akira. Simulations of galaxy mergers constrained by the data (see Methods) suggest that the cool component is part of a tidal stream and is arcing towards the observer from the far West (blueshifted; Fig. 1e) before plunging back towards Akira's centre (redshifted; Fig. 1e).

Previous work has noted similar objects^{20,23–25} but has typically attributed their gaseous dynamics and unusual emission line features to accreted, rotating disks²⁶. However, using a tight constraint on the total gravitational potential derived from the stellar kinematics, we find that the observed second velocity moments, V_{rms} —defined as $V_{\text{rms}} \equiv \sqrt{V^2 + \sigma^2}$, where V is the velocity and σ is the velocity dispersion—of the ionized gas in Akira are far too high to be consistent with motions under the influence of gravity alone (Fig. 1j; see Methods). Regardless of gas inclination or the degree of pressure support, we can rule out any kind of axisymmetric orbital distribution. Perturbations or torques from disk ‘settling’ are also very unlikely to drive discrepancies

Table 1 | Galaxy properties

MaNGA name	MaNGA-ID	RA (J2000.0 deg.)	Dec. (J2000.0 deg.)	z^*	$\log[M_* (M_\odot)]^\dagger$	NUV – r^\ddagger	$\log[\text{SFR} (M_\odot \text{ yr}^{-1})]^\S$	R_e^\parallel (kpc)	$\log[M_h (M_\odot)]^\P$
Akira (host)	1-217022	136.08961	41.48174	0.0244671	10.78	6.1	–4.17	3.88	12.0
Tetsuo (companion)	1-217015	136.11416	41.48621	0.0244647	9.18	3.0	–0.94	1.73	12.0

RA, right ascension; Dec., declination.

*Spectroscopic redshift from NSA catalogue (<http://www.nsatlas.org/data>).

†Stellar mass from MPA-JHU DR7 data release (<http://www.mpa.mpa-garching.mpg.de/SDSS/DR7/Data/stellarmass.html>).

‡Rest-frame NUV – *r* colour from NSA catalogue (<http://www.nsatlas.org/data>).

§Star formation rate (SFR) from SED fitting of SDSS optical and WISE infrared photometry¹⁹; the AGN contribution to the SED is negligible.

||Effective radius from NSA catalogue (<http://www.nsatlas.org/data>).

¶Halo mass, M_h , from a public group catalogue¹⁸.

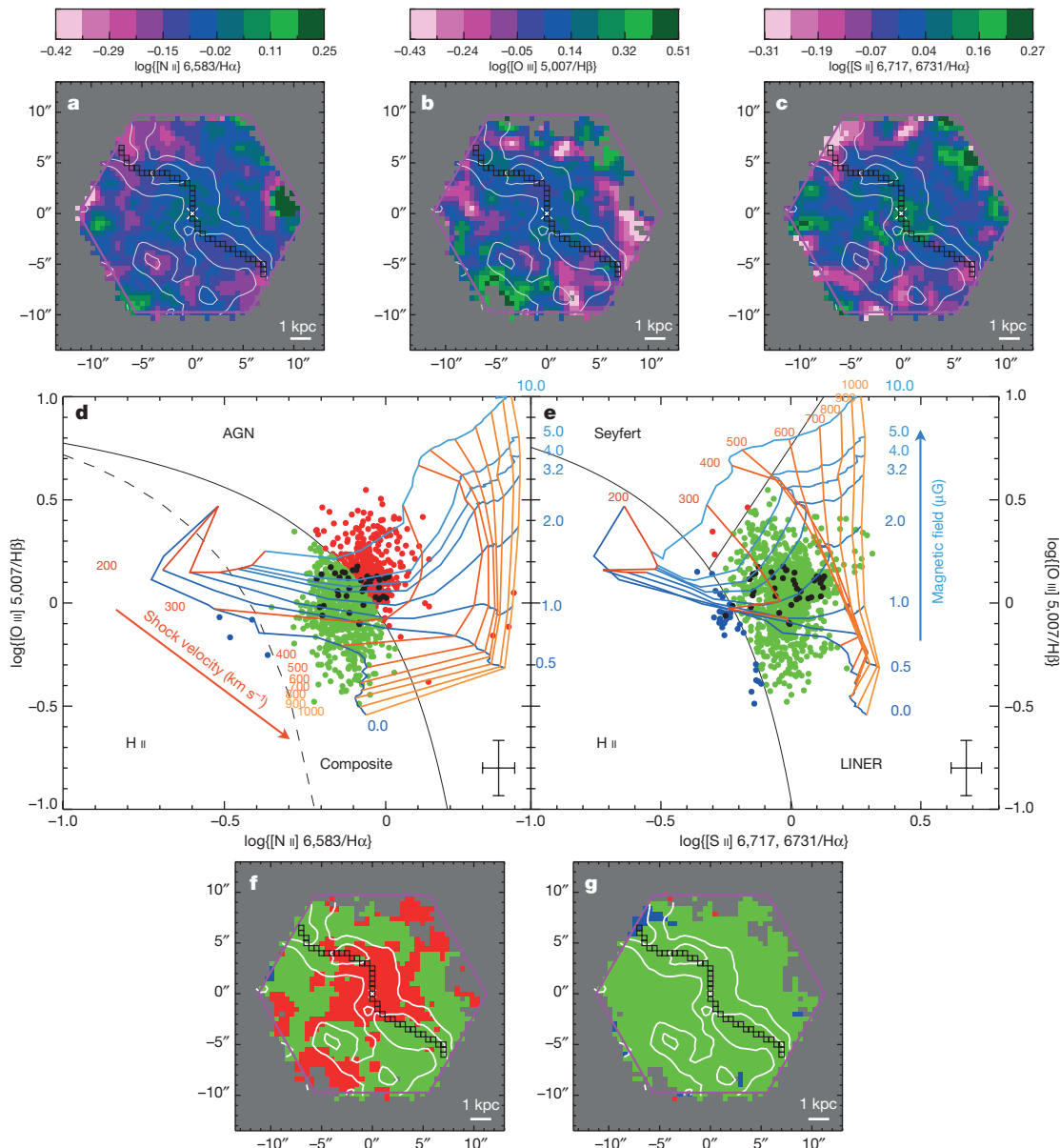
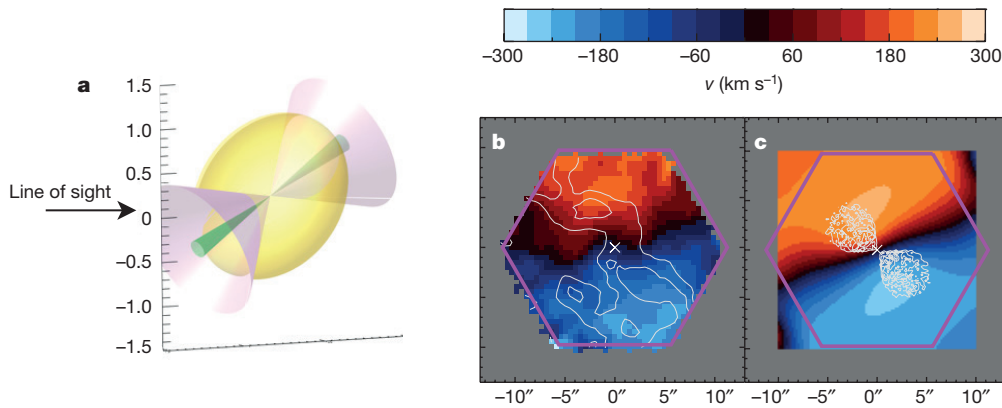


Figure 3 | Diagnostic line-ratio maps of Akira. **a–c**, Maps of line ratios. **a**, $\log\{[\text{N II}] 6,583/\text{H}\alpha\}$, **b**, $\log\{[\text{O III}] 5,007/\text{H}\beta\}$ and **c**, $\log\{[\text{S II}] 6,717, 6,731/\text{H}\alpha\}$, with contours tracing the $\text{H}\alpha$ EW pattern. **d, e**, The $[\text{N II}] 6,583$ and $[\text{S II}] 6,717$, Baldwin-Phillips-Terlevich (BPT) diagrams; error bars represent the 1σ measurement errors propagated to the log line ratios. The solid and dashed lines separate the H II (blue points), AGN (red points), Seyfert (red points), Composite (green points), and LINER

(green points) classifications. Overplotted (orange curves labelled with shock velocity in km s^{-1}) are shock models²⁷, and the black points correspond to the spaxels highlighted by black boxes in the other panels. **f, g**, The resolved $[\text{N II}] 6,583$ and $[\text{S II}] 6,717, 6,731$ BPT maps, that is, each spaxel is coloured by its location on their respective BPT diagram, with contours tracing the $\text{H}\alpha$ EW pattern.

that reach as high as $\sim 100 \text{ km s}^{-1}$. We can express the dynamical inconsistency of the disk hypothesis another way. If we assume such a disk were inclined at $i = 50^\circ$ (see Methods), we estimate that 15%–20% of the disk would be moving at velocities sufficient to escape the galaxy. With similar velocity properties observed for the rest of this new class of galaxies, the disk interpretation also fails to explain why the bisymmetric H α features are always in rough alignment with the major kinematic axis. If arising from internal structure in a moderately face-on disk, this H α EW structure should be randomly oriented compared to the kinematic axis, which is instead determined by the observer's viewing angle.

A relatively simple wind model with a constant radially-outward velocity of 310 km s^{-1} confined to a wide-angle (θ) cone ($2\theta = 80^\circ$) reproduces several qualitative features of the data (Fig. 2; see Methods). The model captures the overall shape of the ionized-gas velocity field and associates the extended (horizontal) zones of high ionized-gas velocity dispersion along the kinematic minor axis with the overlapping projection of approaching and receding surfaces of the inclined wind cone. By assigning somewhat greater wind densities to the cone centre, we can explain the offsets between the projected kinematic major axis of the ionized gas and both the stellar position angle and the H α flux orientation. Furthermore, the bisymmetric H α EW features can be explained by enhanced gas over-densities or shock ionization along the central wind axis. Indeed, Fig. 3d, e demonstrates that line ratios in the H α EW feature (black points and boxes throughout Fig. 3) tend to cluster and are consistent with those predicted by 'fast' shock models²⁷ with velocities of $200\text{--}400 \text{ km s}^{-1}$.

The wind's driving mechanism probably originates in Akira's active (at radio wavelengths) galactic nucleus (AGN), which is detected in FIRST (Faint Images of the Radio Sky at Twenty-Centimeters) data with a luminosity density of $L_{1.4 \text{ GHz}} = 1.6 \times 10^{21} \text{ W Hz}^{-1}$, and is most consistent with being a point source according to higher-resolution ($1.5''$) follow-up Jansky VLA (Very Large Array) radio observations (W.R., manuscript in preparation). Since this AGN lacks obvious extended radio jets, the feedback of this AGN is most likely to manifest in small-scale jets ($< 1 \text{ kpc}$) or uncollimated winds^{28,29}. Despite an Eddington ratio of $\lambda = 3.9 \times 10^{-4}$, energetics arguments show that the AGN's mechanical output ($P_{\text{mech}} = 8.1 \times 10^{41} \text{ erg s}^{-1}$) is sufficient to supply the wind's kinetic power ($\dot{E}_{\text{wind}} \approx 10^{39} \text{ erg s}^{-1}$; see Methods). Moreover, the wind can inject sufficient energy, coupled to the ambient gas through the turbulent dynamics observed (Figs 1i, 3a–c), to balance the cooling rate (\dot{E}_{gas}) in both the ionized and cool gas ($\dot{E}_{\text{gas}} \approx 10^{39} \text{ erg s}^{-1}$). Indeed, the amount of cool Na D gas ($M_{\text{cool gas}} \approx 10^8 M_\odot$) implies a star formation rate of $\text{SFR} \approx 1 \times 10^{-2} M_\odot \text{ yr}^{-1}$, which is much higher than the estimated¹⁹ $\text{SFR}_{\text{Akira}} = 7 \times 10^{-5} M_\odot \text{ yr}^{-1}$ that is derived from well-detected WISE photometry. The picture that emerges is one in which cool gas inflow to Akira, triggered by the interaction with Tetsuo, has initiated a relatively low-power AGN-driven wind that is nonetheless able to heat the surrounding gas through turbulence and shocks and thereby prevent any substantial star formation.

As with Akira, the other galaxies in this class show little or no ongoing star formation, and the majority harbour similarly weak radio point sources (according to follow-up Jansky VLA observations) that would be classified as 'jet mode', 'kinetic mode' or 'radio mode' AGN^{17,15}. With similar levels of fast-moving ionized gas oriented along enhanced ionized emission, we conclude that AGN-driven winds are present in these systems as well, and represent an important heating source. Because the full spatial extent of these winds may exceed the field-of-view of our observations, a lower limit of $\sim 10^7 \text{ yr}$ for the timescale of this phenomenon is given by the radial extent divided by the typical wind velocity. Assuming all quiescent galaxies experience these AGN-driven winds, the $\sim 5\%$ occurrence rate (averaged over the full mass range) implies an episodic behaviour that leads us to name these objects 'red geysers'. Present primarily below $M_* \lesssim 10^{11} M_\odot$, these galaxies lie in isolated haloes with moderate masses¹⁸ ($M_{\text{halo}} \approx 10^{12} M_\odot$) and exhibit no signs

of major interactions. Their implied trigger rate (at most, a few episodes per Gyr) may be related to minor mergers (approximately one per Gyr; ref. 30) as well as central accretion of ambient hot gas from stellar mass loss⁸. These red geysers may exemplify how typical quiescent galaxies maintain their quiescence.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 October 2015; accepted 30 March 2016.

1. Bell, E. F. *et al.* Nearly 5000 distant early-type galaxies in COMBO-17: a red sequence and its evolution since $z \sim 1$. *Astrophys. J.* **608**, 752–767 (2004).
2. Bundy, K. *et al.* The mass assembly history of field galaxies: detection of an evolving mass limit for star-forming galaxies. *Astrophys. J.* **651**, 120–141 (2006).
3. Faber, S. M. *et al.* Galaxy luminosity functions to $z \sim 1$ from DEEP2 and COMBO-17: implications for red galaxy formation. *Astrophys. J.* **665**, 265–294 (2007).
4. Ilbert, O. *et al.* Galaxy stellar mass assembly between $0.2 < z < 2$ from the S-COSMO survey. *Astrophys. J.* **709**, 644–663 (2010).
5. Di Matteo, T., Springel, V. & Hernquist, L. Energy input from quasars regulates the growth and activity of black holes and their host galaxies. *Nature* **433**, 604–607 (2005).
6. Hopkins, P. F. *et al.* A unified, merger-driven model of the origin of starbursts, quasars, the cosmic X-ray background, supermassive black holes, and galaxy spheroids. *Astrophys. J.* **163**, 1–49 (2006).
7. Heckman, T. M. & Best, P. N. The coevolution of galaxies and supermassive black holes: insights from surveys of the contemporary universe. *Annu. Rev. Astron. Astrophys.* **52**, 589–660 (2014).
8. Ciotti, L. & Ostriker, J. P. Cooling flows and quasars: different aspects of the same phenomenon? I. Concepts. *Astrophys. J.* **487**, L105–L108 (1997).
9. Benson, A. J. *et al.* What shapes the luminosity function of galaxies? *Astrophys. J.* **599**, 38–49 (2003).
10. Booth, C. M. & Schaye, J. The interaction between feedback from active galactic nuclei and supernovae. *Sci. Rep.* **3**, 1738 (2013).
11. Croton, D. J. *et al.* The many lives of active galactic nuclei: cooling flows, black holes and the luminosities and colours of galaxies. *Mon. Not. R. Astron. Soc.* **365**, 11–28 (2006).
12. Bower, R. G. *et al.* Breaking the hierarchy of galaxy formation. *Mon. Not. R. Astron. Soc.* **370**, 645–655 (2006).
13. Ciotti, L., Ostriker, J. P. & Proga, D. Feedback from central black holes in elliptical galaxies. III. Models with both radiative and mechanical feedback. *Astrophys. J.* **717**, 708–723 (2010).
14. Fabian, A. C. A very deep Chandra observation of the Perseus cluster: shocks, ripples and conduction. *Mon. Not. R. Astron. Soc.* **366**, 417–428 (2006).
15. Fabian, A. C. Observational evidence of active galactic nuclei feedback. *Annu. Rev. Astron. Astrophys.* **50**, 455–489 (2012).
16. Lin, Y.-T. & Mohr, J. J. Radio sources in galaxy clusters: radial distribution, and 1.4 GHz and K-band bivariate luminosity function. *Astrophys. J. Suppl. Ser.* **170**, 71–94 (2007).
17. Bundy, K. *et al.* Overview of the SDSS-IV MaNGA survey: mapping nearby galaxies at Apache Point observatory. *Astrophys. J.* **798**, 7 (2015).
18. Yang, X. *et al.* Galaxy groups in the SDSS DR4. I. The catalog and basic properties. *Astrophys. J.* **671**, 153–170 (2007).
19. Chang, Y.-Y., van der Wel, A., da Cunha, E. & Rix, H.-W. Stellar masses and star formation rates for 1M galaxies from SDSS+WISE. *Astrophys. J. Suppl. Ser.* **219**, 8 (2015).
20. Sarzi, M. *et al.* The SAURON project — XVI. On the sources of ionization for the gas in elliptical and lenticular galaxies. *Mon. Not. R. Astron. Soc.* **402**, 2187–2210 (2010).
21. Yan, R. & Blanton, M. R. The nature of LINER-like emission in red galaxies. *Astrophys. J.* **747**, 61 (2012).
22. Belfiore, F. *et al.* P-MaNGA galaxies: emission-lines properties — the gas ionization and chemical abundances from prototype observations. *Mon. Not. R. Astron. Soc.* **449**, 867–900 (2015).
23. Kehrig, C. *et al.* The ionized gas in the CALIFA early-type galaxies. *Astron. Astrophys.* **540**, A11 (2012).
24. Allen, J. T. *et al.* The SAMI galaxy survey: unveiling the nature of kinematically offset active galactic nuclei. *Mon. Not. R. Astron. Soc.* **451**, 2780–2792 (2015).
25. Gomes, J. M. *et al.* The warm ionized gas in CALIFA early-type galaxies: 2D emission-line patterns and kinematics for 32 galaxies. *Astron. Astrophys.* **588**, A68 (2016).
26. Lagos, C. P. *et al.* The origin of the atomic and molecular gas contents of early-type galaxies — II. Misaligned gas accretion. *Mon. Not. R. Astron. Soc.* **448**, 1271–1287 (2015).
27. Allen, M. G., Groves, B. A., Dopita, M. A., Sutherland, R. S. & Kewley, L. J. The MAPPINGS III library of fast radiative shock models. *Astrophys. J. Suppl. Ser.* **178**, 20–55 (2008).
28. Ostriker, J. P., Choi, E., Ciotti, L., Novack, G. S. & Proga, D. Momentum driving: which physical processes dominate active galactic nucleus feedback? *Astrophys. J.* **722**, 642–652 (2010).

29. Yuan, F. & Narayan, R. Hot accretions flows around black holes. *Annu. Rev. Astron. Astrophys.* **52**, 529–588 (2014).
30. Hopkins, P. F. *et al.* Mergers and bulge formation in Λ CDM: which mergers matter? *Astrophys. J.* **715**, 202–229 (2010).

Acknowledgements We are grateful to Y.-Y. Chang for checks on the SED fitting and implied SFR. We thank S. Juneau, J. Newman, H. Fu, K. Nyland, and S. F. Sánchez for discussions and comments. This work was supported by the World Premier International Research Center Initiative (WPI Initiative), MEXT, Japan, and JSPS KAKENHI grant no. 15K17603. A.W. acknowledges support of a Leverhulme Trust Early Career Fellowship. S.P. acknowledges support from the Japan Society for the Promotion of Science (JSPS long-term invitation fellowship). M.C. acknowledges support from a Royal Society University Research Fellowship. W.R. is supported by a CUUniverse Grant (CUAASC) from Chulalongkorn University. Funding for the Sloan Digital Sky Survey IV (SDSS-IV) has been provided by the Alfred P. Sloan Foundation, the US Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS website is www.sdss.org. SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration, including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, the Chilean Participation Group, the French Participation Group, Harvard-Smithsonian Center for Astrophysics, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU)/University of Tokyo, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-

Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatory of China, New Mexico State University, New York University, University of Notre Dame, Observatório Nacional/MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, UK Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University and Yale University.

Author Contributions E.C. and K.B. discovered the described sources, interpreted the observations, built the wind model, and wrote the manuscript. M.C. constructed dynamical models. S.P. carried out numerical merger simulations to model the data. W.R. obtained and reduced the JVLA data. K.W. fitted disk models. K.B., R.Y., M.B., N.D., D.R.L., D.A.W., K.Z., A.W., K.L.M. and D.T. contributed to the design and execution of the survey. F.B. provided initial velocity and line-ratio maps. B.V. provided the modelled extinction map. Y.C. and K.R. contributed to the Na D interpretation. All authors contributed to the interpretation of the observations and the writing of the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.C. (ec2250@gmail.com).

METHODS

Observations. The data used in this work come from the ongoing MaNGA survey^{17,31,32} using the SDSS 2.5-metre telescope³³. One of three programs comprising SDSS-IV, MaNGA is obtaining spatially resolved spectroscopy for 10,000 nearby galaxies with $\log(M_*/M_\odot) \gtrsim 9$ and a median redshift of $z \approx 0.04$. The r-band signal-to-noise ratio (S/N) in the galaxy outskirts is 4–8 Å⁻¹, and the wavelength coverage is 3,600–10,300 Å. The effective spatial resolution is 2.4'' (full width at half maximum; FWHM) with an instrumental spectral resolution of ≈ 60 km s⁻¹. The sample and data products used here were drawn from the internal MaNGA Product Launch-3 (MPL-3), which includes ~ 700 galaxies observed before April 2015 and will be publicly available in the thirteenth SDSS data release.

Ancillary data are from the NASA-Sloan Atlas (NSA³⁴), MPA-JHU DR7 data release³⁵, and other recent works^{18,19}. We assume a flat cosmological model with $H_0 = 70$ km s⁻¹ Mpc⁻¹, $\Omega_m = 0.30$ and $\Omega_\Lambda = 0.70$, and all magnitudes are given in the AB magnitude system³⁶.

The Data Analysis Pipeline (DAP), which uses pPXF³⁷ and the MIUSCAT stellar library³⁸, fits the stellar continuum in each spaxel and produces estimates of the stellar kinematics. Flux and EW measurements were measured through simple flux-summing³⁹ after we subtract the stellar continuum. We only show flux and EW measurements with $S/N > 3$ Å⁻¹ in the wavelength range around a given line. Ionized gas kinematics, that is, velocity ($\nu_{\text{ionized gas}}$) and velocity dispersion ($\sigma_{\text{ionized gas}}$), were estimated by fitting a single Gaussian to the H α emission line.

W_{80} . W_{80} is a non-parametric measure of line widths; it is defined to contain 80% of the emission-line flux⁴⁰.

Na D measurements. Using a spectral fitting code^{41,42}, we present the dust extinction map of Akira in Extended Data Fig. 1. The superimposed Na D contours (from Fig. 1d) overlap with enhanced extinction (darker spaxels), supporting the association of the offset Na D absorption with cool foreground material.

To measure the line-of-sight (LOS) velocity of this Na D-absorbing material, which we defined as spaxels with Na D EW > 3.5 Å, we first subtract the stellar continuum fit determined for Akira by the DAP. Extended Data Fig. 2a–c shows the stellar continuum fits (red) around the Na D doublet (the two black vertical lines mark the expected locations of the Na D doublet) for three spectra. Extended Data Fig. 2a shows data from a recent work⁴³, Extended Data Fig. 2b shows the central spaxel of Akira, marked by the “x” in Fig. 1d, e and Extended Data Fig. 1, and Extended Data Fig. 2c shows a spaxel to the northwest of Akira, marked by the single box in the upper right of Fig. 1d, e and Extended Data Fig. 1. We then examine the residual absorption as a function of wavelength, as shown in Extended Data Fig. 2d–f.

Focusing on Extended Data Fig. 2d–f, we determine the line centroids in Akira by first defining a reference Na D profile for typical, cold interstellar medium gas at rest. We use the stacked, continuum-subtracted spectrum of Na D from a large set of highly inclined disk galaxies from this reference⁴³, which is shown in the left panel. We define an at-rest line centroid for cool Na D gas by averaging the wavelengths in this profile, each weighted by the amplitude of the residual absorption at that wavelength (weighting is performed within the green region). The resulting centroid is marked by the dotted grey vertical line, which is repeated in the two right panels for reference. In the same way, we determine line centroids for the observed residual profiles across the Na D-absorbing material in Akira, which is marked by the blue vertical lines in the two right panels. We then calculate the velocity difference between the reference Na D centroid and the observed Na D centroid in these spaxels of Akira; this velocity difference is shown in the upper left in Extended Data Fig. 2e, f.

Merger simulations. We modelled the interaction between Akira and Tetsuo using the GADGET-2⁴⁴ code and the methodology described in a recent work⁴⁵. These simulations are constrained by the available data and contain more than four million particles that account for stars, dark matter, and gas (we only consider gas in Tetsuo). These simulations also include cooling, star formation, and supernova feedback, but not AGN feedback nor the proposed wind. The initial total mass merger ratio is $\sim 1:10$, but because Tetsuo loses mass during the interaction, this ratio falls to $\sim 1:20$ at the time most closely matching the observations (the observed stellar mass merger ratio is 1:40). According to the best-matching viewing angle for this prograde encounter, Tetsuo starts in the foreground to the lower-right of Akira and begins arcing over the top and away from the observer (see Extended Data Fig. 3a–d). After a glancing blow with Akira, a tidal bridge is generated that loops back and passes through Akira to form the shell structure seen to the lower right (Extended Data Fig. 3d). This snapshot at $t = 0.56$ Gyr best matches the SDSS r image (Extended Data Fig. 3f), and it indicates that Tetsuo is behind Akira. Extended Data Fig. 3e shows a composite stars+gas representation at this snapshot; it indicates that a stream of cool gas from Tetsuo has followed the stellar bridge that is behind Akira, penetrated close to Akira's centre, emerged in front of Akira on its lower-right side, and approaches the observer.

The shape of the tidal bridge and shell to the southwest in the SDSS image (Extended Data Fig. 3f) provides the most significant constraints on the simulation and its viewing angle. An important cross-check is that the orientation of Tetsuo's stellar and ionized gas velocity fields (also observed by MaNGA) are reproduced as well. The geometry and velocity scale of the cool gas is similar to the observed Na D component (Fig. 1d, e), but there are differences from the observations. Portions of the observed Na D gas appear to be falling back into Akira (redshift; Fig. 1e), but these are not seen in the simulation until a later time step. The observed cool gas orientation is also more horizontal while the simulation predicts the gas stream stretches further (Extended Data Fig. 3e). But we emphasize that we only detect cool gas in absorption where there are background stars from host galaxy, whereas the simulation allows us to see the full extent of the cool gas. Differences between the simulations and observations may also arise from inaccuracies in the initialization of the merger simulation (mass ratios, gas mass fractions, angular momentum alignment, and so on), limitations in the hydrodynamic gas treatment, or missing components in the simulation such as Akira's gas supply and the proposed AGN-driven wind.

Dynamical modelling evidence against the presence of disks. Jeans Anisotropic Modelling (JAM⁴⁶), which uses the Multi-Gaussian Expansion (MGE^{47,48}) parametrization for mass and light distributions, was performed on Akira and other red geysers to model their stellar kinematics and gravitational potential. The JAM model derives a 3D stellar density by de-projecting the observed SDSS r-band photometry using an MGE fit. The modelled potential includes an NFW⁴⁹ dark matter halo. The JAM model has four free parameters: the inclination i , anisotropy β_z , stellar M/L (that is, mass-to-light ratio) and halo mass. These are optimized by fitting the model prediction for the second velocity moments, $V_{\text{rms}} \equiv \sqrt{V^2 + \sigma^2}$, to the observed MaNGA stellar kinematics. Through a number of systematics tests, we find that the best-fit stellar inclination is $i = 41^\circ$, with an upper limit of $i = 50^\circ$. Although there is some covariance between the model parameters, the resulting total mass profile is extremely robust⁵⁰.

With the total gravitational potential defined from the stellar JAM modelling above, we can predict projected second velocity moment (V_{rms}) maps of gas under the assumption of axisymmetric orbital distributions. We treat the H α -emitting gas clouds as a ‘tracer’ population of the underlying potential. Its flux distribution is modelled by a separate MGE (distinct from the stellar component) enabling de-projection of the observed H α surface brightness. The Jeans equations are then solved for this tracer, within the fixed potential, to predict the V_{rms} allowed by the given mass distribution. We emphasize that the second moments are independent of the degree of circular motion versus ‘random’ motion in the hypothesized disk. The analysis does not account for non-gravitational drivers of turbulent pressure, such as from the AGN-driven wind we propose. In Extended Data Fig. 4 (see also Fig. 1j) we show results for gaseous inclinations of $i = 46^\circ$ (the minimum allowed by the $b/a = 0.7$ from GALFIT fits of the H α flux, corresponding to an intrinsic axis ratio $q = 0.12$; see below) and the most extreme case of $i = 90^\circ$ (an edge-on axisymmetric density). In either case, the allowed V_{rms} is far below the observed V_{rms} .

With discrepancies as high as ~ 100 km s⁻¹, torques of the same order as the gravitational potential itself would be required to explain the data, making a ‘disturbed’ disk a highly unlikely explanation. It is possible to imagine a very chaotic accretion scenario where the JAM assumptions of axisymmetry and stability completely break down, although in this case an ordered ionized-gas velocity gradient of the kind observed seems unlikely. Such a scenario would also struggle to explain how the high ionized-gas velocity dispersions are generated and why enhanced H α flux is observed along the gradient in the ionized-gas velocity field. Similarly, because line widths of $W_{80} \approx 500$ km s⁻¹ could not be sustained by accreting tidal streams or caused by tidal torques, multiple overlapping gas streams would have to conspire to produce the widespread high velocity dispersion observed (Fig. 1i) while maintaining an ordered velocity gradient pattern. A similar set of coincidences would be required for each galaxy in the rest of the red geyser sample.

Not surprisingly, tilted-disk models⁵¹ that fit the ionized velocity field alone do a poor job for the red geyser sample. Characterizing the goodness-of-fit by an error-weighted average residual, the majority of red geysers exhibit residuals that place them among the worst 5% of fitted MaNGA galaxies with ‘disk-like’ kinematics. Here, disk-like refers to galaxies with reasonable agreement between stellar and gaseous systemic velocities, dynamical centres, position angles, and inclinations.

Finally, we use the dynamically constrained potential to estimate a local escape velocity and compare this to the inferred velocity distribution of a putative disk. Several assumptions are required, but the results are informative. We obtain a rough estimate of escape velocity, $\nu_{\text{esc}} \approx 400 \pm 50$ km s⁻¹, by integrating the potential from a projected radius of $7''$ (3.4 kpc or just under $1 R_e$) to $4 R_e$ (16 kpc) and assuming a gentle decline in the circular velocity at large radius. We then use GALFIT⁵² to model the observed H α flux surface brightness, finding a consistent projected axis ratio of $b/a = 0.7 \pm 0.02$, regardless of the assumed model profile (exponential, de Vaucouleurs, or free Sérsic) and despite significant structure in the residuals

(of the order of $\sim 10\%$ – 15%). Hypothesizing a disk with an intrinsic axis ratio, $q = 0.4$, roughly twice as 'fat' as typical disks⁵³, we estimate an inclination of $i = 50^\circ$. This is also the upper limit of inclinations allowed for the stellar kinematics, and precession should align accreted material with the stellar distribution in roughly a few dynamical times⁵⁴ (unless there is a source of incoming misaligned gas⁵⁵). We de-project the observed mean velocities using this inclination and consider the distribution of velocities about this mean. Roughly 15%–20% of the gas, that is, with velocities greater than 1σ from the mean, would exceed the escape velocity under these assumptions.

Wind model. We construct a simple wind model that reproduces many qualitative features of the MaNGA observations. In this model, the wind assumes a wide-angle biconical form centred on the galaxy nucleus. Within the bicone, the wind has a constant amplitude, radially-outward velocity⁵⁶. We assume that warm gas clouds entrained by the wind trace this velocity structure and emit flux in strong emission lines primarily in response to the local ionization field supplied by the stars^{20–22}. The projected wind velocity field to first order is therefore a convolution of the wind geometry with the galaxy's 3D luminosity profile.

To realize the model, we populate a randomized 3D Cartesian grid of points with the galaxy at the centre and assign each point a weight equal the value of an axisymmetric Hernquist density profile sampled at that point⁵⁷. This density profile is fixed to reproduce the imaging and JAM constraints on the stellar component, namely an intrinsic (3D) axis ratio of 0.4, an inclination of 41° , a projected major-axis effective radius of $R_e \approx 7''$, and an on-sky PA of 53° . For a given wind opening angle and inclination, we weight the projected line-of-sight component of the wind velocity at each point inside the bicone by its Hernquist profile value. Projected quantities are smoothed to the spatial resolution of the MaNGA data ($2.4''$, FWHM). To model a potential enhancement of gas densities or shocks along the central axis of the bicone, we implement a second set of weights defined with respect to the bicone that decrease exponentially (with a variable characteristic angle) as a function of the angular distance from the bicone's axis.

By experimenting with different choices for the wind's opening angle, inclination, length, intrinsic velocity, (and central weighting, if desired), we explored possible wind model solutions. Most have opening angles of $2\theta \approx 80^\circ$ and steep inclinations ($\sim 70^\circ$) towards the line-of-sight. One example is shown in Fig. 2. This wind model has an opening angle of $2\theta = 80^\circ$, an inclination of 75° , PA = 55° , and a length of $2R_e$. We have assumed a constant radially outward velocity within the wind of $v_{\text{wind}} = 310 \text{ km s}^{-1}$. We associate the observed, bisymmetric regions of enhanced H α (white contours on the observed velocity field; Fig. 2b) with the wind's central axis. The projection of this $\pm 10''$ region is overplotted with white contours on the modelled velocity field (Fig. 2c). The wind density is assumed to decline as an exponential function of the angular distance with a characteristic angle of $\alpha = 10^\circ$.

Shock models. Shock models²⁷ with twice the solar atomic abundances, shock velocities of 200–400 km s^{-1} , magnetic fields of 0.5–10 μG , and preshock densities of unity, were used in Fig. 3.

Inferring the presence of an AGN in Akira. The presence of a central radio source and the absence of star formation in Akira imply the presence of an AGN. Quantitatively, we can confirm the presence of an AGN by comparing the expected SFR inferred from the radio luminosity of Akira to the estimated SFR from SED fitting of SDSS and WISE photometry¹⁹. We first calculate the radio luminosity density of Akira using $L_{1.4\text{GHz}} = 4\pi d_L^2 F_{1.4\text{GHz}}$, where $F_{1.4\text{GHz}}$ is the integrated flux density (of 1.2 mJy) from FIRST⁵⁸, and d_L is the luminosity distance. This calculation yields $L_{1.4\text{GHz}} = 1.6 \times 10^{21} \text{ W Hz}^{-1}$. Using the radio SFR calibration⁵⁹, we infer $\text{SFR} = 1 M_\odot \text{ yr}^{-1}$. This level of star formation in Akira is ruled out at more than 97.5% confidence¹⁹, indicating that the most likely source of this radio emission is an AGN.

Eddington ratio and AGN power. To calculate the Eddington ratio (λ) of Akira, we use the Eddington-scaled accretion rate⁶⁰, which is more applicable to radio-detected AGN: $\lambda = (L_{\text{rad}} + L_{\text{mech}})/L_{\text{Edd}}$, where L_{rad} is the bolometric radiative luminosity, L_{mech} is the jet mechanical luminosity, and L_{Edd} is the Eddington limit. To calculate L_{rad} , we converted the [O III] 5,007 Å flux from the central $2''$ ($\sim 1 \text{ kpc}$) radius aperture of Akira, $F_{[\text{O III}]}$, to a luminosity: $L_{[\text{O III}]} = 4\pi d_L^2 F_{[\text{O III}]}$ = $1.7 \times 10^{39} \text{ erg s}^{-1}$. Even though the central [O III] 5007 flux is probably not entirely due to AGN photoionization (evolved stars and shocks probably contribute), for this order-of-magnitude calculation we will make the simplifying assumption that it does. Using the relation⁶¹ $L_{\text{rad}} = 3,500 L_{[\text{O III}]}$, we obtain $L_{\text{rad}} = 5.9 \times 10^{42} \text{ erg s}^{-1}$.

Acknowledging that Akira is in a lower mass and energy output regime than those in which expanding X-ray bubbles have been observed, we nonetheless applied the following relation⁶² to calculate the jet mechanical luminosity: $L_{\text{mech}} = 7.3 \times 10^{36} (L_{1.4\text{GHz}}/10^{24} \text{ W Hz}^{-1})^{0.70} \text{ W}$, which results in $L_{\text{mech}} = 8.1 \times 10^{34} \text{ W} = 8.1 \times 10^{41} \text{ erg s}^{-1}$.

Finally, to calculate L_{Edd} , we first estimate the black hole mass, M_{BH} , using the relation⁶³ $\log(M_{\text{BH}}/M_\odot) = 8.32 + 5.64 \log[\sigma_{\text{star}}/(200 \text{ km s}^{-1})]$, with $\sigma_{\text{star}} = 185.5 \text{ km s}^{-1}$ from the central $2''$ radius aperture, yielding $\log(M_{\text{BH}}/M_\odot) = 8.1$. We calculate the classical Eddington limit with $L_{\text{Edd}} = 3.3 \times 10^4 M_{\text{BH}} = 4.5 \times 10^{12} L_\odot = 1.7 \times 10^{46} \text{ erg s}^{-1}$.

Inserting these numbers into $\lambda = (L_{\text{rad}} + L_{\text{mech}})/L_{\text{Edd}}$ yields $\lambda = 3.9 \times 10^{-4}$, suggesting that the accretion onto this black hole is at a low rate and/or radiatively inefficient; these types of AGN have been termed low-energy, kinetic mode, jet mode, or radio mode AGN^{7,15,60}.

Ionized gas energetics. Assuming warm ionized gas clouds with a temperature of 10^4 K and using the observed [S II] ratio, we estimate⁶⁴ an electron density, n_e , of 100 cm^{-3} . With this value of n_e , we estimate⁶⁵ the lower limits on the ionized gas mass from the H α line flux, $M_{\text{warm, H}\alpha} \approx 6 \times 10^5 M_\odot$. We can derive similar estimates⁶⁶ based on the H β and [O III] flux, obtaining $M_{\text{warm, H}\beta} \approx 4 \times 10^5 M_\odot$ and $M_{\text{warm, [O III]}} \approx 2 \times 10^4 M_\odot$. We adopt an approximate $M_{\text{warm}} \approx 10^5 M_\odot$.

To approximate the energy associated with a wind driving the observed velocities in the ionized gas, we adopt the kinetic energy⁴⁰, $E_{\text{wind}} \approx 0.5 M_{\text{warm}} v_{\text{wind}}^2$, with $v_{\text{wind}} = 300 \text{ km s}^{-1}$. To estimate the wind power, we divide E_{wind} by the characteristic wind timescale of 10^7 yr , derived by dividing Akira's optical radius (the observed extent of the wind) by v_{wind} . We obtain $\dot{E}_{\text{wind}} \approx 10^{39} \text{ erg s}^{-1}$. Because the ionized gas mass is probably a lower limit, \dot{E}_{wind} is likely to be an underestimate. The gas cooling rate is estimated using a method from the literature⁶⁷.

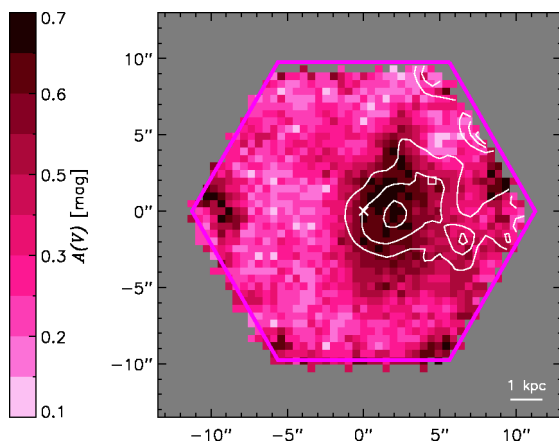
Star formation in the Na D cool gas. To estimate the expected star formation rate associated with the cool Na D gas, we first estimate⁶⁸ the total hydrogen column density ($N_{\text{HI}} + 2N_{\text{H}_2}$), where N is the column density, from the dust extinction presented in Extended Data Fig. 1. Integrating over the $\sim 4 \text{ kpc}^2$ region of enhanced extinction, we find a total gas mass of $M_{\text{cool}} \approx 10^8 M_\odot$ or a surface mass density of $\Sigma_{\text{cool}} \approx 3 \times 10^7 M_\odot \text{ kpc}^{-2}$. To apply the Kennicutt relation⁶⁹, we first account for fact that the Na D material is unlikely to be distributed in a thin, face-on disk. Assuming the Kennicutt relation holds with respect to volumetric density, we scale Σ_{cool} by the ratio of scale heights between a typical star-forming spiral ($H_{\text{Kennicutt}} \approx 0.6 \text{ kpc}$; ref. 70) and an estimate for the Na D material's scale height, H_{NaD} . We set H_{NaD} to $\sim 3 \text{ kpc}$, which is approximately the effective radius (R_e) of Akira. These assumptions yield $\text{SFR} \approx 10^{-2} M_\odot \text{ yr}^{-1}$, roughly 100 times higher than the estimate for Akira ($\text{SFR}_{\text{Akira}} = 7 \times 10^{-5} M_\odot \text{ yr}^{-1}$)¹⁹.

Sample size. No statistical methods were used to predetermine sample size.

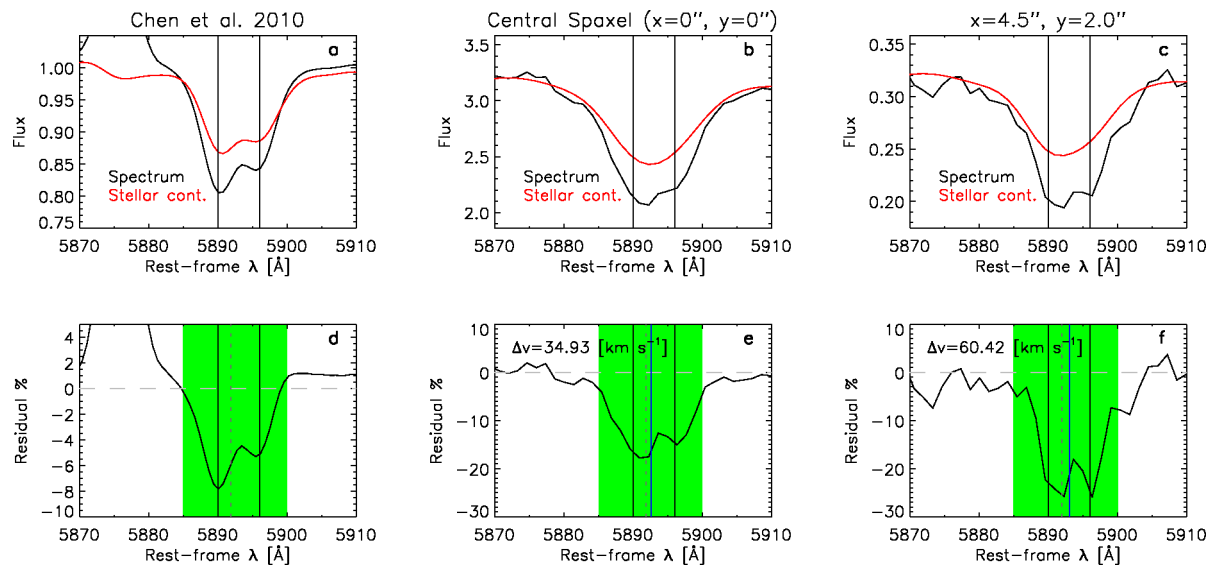
Code availability. The JAM code is available at <http://www-astro.physics.ox.ac.uk/~mxc/software/#jam>

- Drory, N. *et al.* The MaNGA integral field unit fiber feed system for the Sloan 2.5 m telescope. *Astron. J.* **149**, 77 (2015).
- Law, D. R. *et al.* Observing strategy for the SDSS-IV/MaNGA IFU galaxy survey. *Astron. J.* **150**, 19 (2015).
- Gunn, J. E. *et al.* The 2.5 m telescope of the Sloan Digital Sky Survey. *Astron. J.* **131**, 2332–2359 (2006).
- Blanton, M. <http://www.nsatlas.org/data> (2009; accessed 30 April 2016).
- The MPA-JHU DR7 release of spectrum measurements. <http://www.mpa-garching.mpg.de/SDSS/DR7/Data/stellarmass.html> (2010; accessed 30 April 2016).
- Oke, J. B. & Gunn, J. E. Secondary standard stars for absolute spectrophotometry. *Astrophys. J.* **266**, 713–717 (1983).
- Cappellari, M. & Ermsellern, E. Parametric recovery of line-of-sight velocity distributions from absorption-line spectra of galaxies via penalized likelihood. *Publ. Astron. Soc. Pacif.* **116**, 138–147 (2004).
- Vazdekis, A. *et al.* MILES: extended MILES spectral coverage — I. Stellar population synthesis models. *Mon. Not. R. Astron. Soc.* **424**, 157–171 (2012).
- Yan, R. *et al.* On the origin of [OII] emission in red-sequence and poststarburst galaxies. *Astrophys. J.* **648**, 281–298 (2006).
- Harrison, C. M., Alexander, D. M., Mullaney, J. R. & Swinbank, A. M. Kiloparsec-scale outflows are prevalent among luminous AGN: outflows and feedback in the context of the overall AGN population. *Mon. Not. R. Astron. Soc.* **441**, 3306–3347 (2014).
- Fritz, J. *et al.* WINGS-SPE II: a catalog of stellar ages and star formation histories, stellar masses and dust extinction values for local clusters galaxies. *Astron. Astrophys.* **526**, A45 (2011).
- Fritz, J. *et al.* WINGS-SPE. III. Equivalent width measurements, spectral properties, and evolution of local cluster galaxies. *Astron. Astrophys.* **566**, A32 (2014).
- Chen, Y.-M. *et al.* Absorption-line probes of the prevalence and properties of outflows in present-day star-forming galaxies. *Astron. J.* **140**, 445–461 (2010).
- Springel, V. The cosmological simulation code GADGET-2. *Mon. Not. R. Astron. Soc.* **364**, 1105–1134 (2005).
- Peirani, S. *et al.* Composite star formation histories of early-type galaxies from minor mergers: prospects for WFC3. *Mon. Not. R. Astron. Soc.* **405**, 2327–2338 (2010).
- Cappellari, M. Measuring the inclination and mass-to-light ratio of axisymmetric galaxies via anisotropic Jeans models of stellar kinematics. *Mon. Not. R. Astron. Soc.* **390**, 71–86 (2008).

47. Emsellem, E., Monnet, G. & Bacon, R. The multi-Gaussian expansion method: a tool for building realistic photometric and kinematical models of stellar systems I. The formalism. *Astron. Astrophys.* **285**, 723–738 (1994).
48. Cappellari, M. Efficient multi-Gaussian expansion of galaxies. *Mon. Not. R. Astron. Soc.* **333**, 400–410 (2002).
49. Navarro, J. F., Frenk, C. S. & White, S. D. M. The structure of cold dark matter halos. *Astrophys. J.* **462**, 563–575 (1996).
50. Li, H. *et al.* Assessing the Jeans anisotropic multi-Gaussian expansion method with the Illustris simulation. *Mon. Not. R. Astron. Soc.* **455**, 3680–3692 (2016).
51. Andersen, D. R. & Bershady, M. A. The photometric and kinematic structure of face-on disk galaxies. III. Kinematic inclinations from H α velocity fields. *Astrophys. J.* **768**, 41 (2013).
52. Peng, C. Y., Ho, L. C., Impey, C. D. & Rix, H.-W. Detailed structural decomposition of galaxy images. *Astron. J.* **124**, 266–293 (2002).
53. Bershady, M. A. *et al.* Galaxy disks are submaximal. *Astrophys. J.* **739**, L47 (2011).
54. Tohline, J. E., Simonson, G. F. & Caldwell, N. Using gaseous disks to probe the geometric structure of elliptical galaxies. *Astrophys. J.* **252**, 92–101 (1982).
55. van de Voort, F. *et al.* The creation and persistence of a misaligned gas disc in a simulated early-type galaxy. *Mon. Not. R. Astron. Soc.* **451**, 3269–3277 (2015).
56. Bouché, N. *et al.* Physical properties of galactic winds using background quasars. *Mon. Not. R. Astron. Soc.* **426**, 801–815 (2012).
57. Dehnen, W. & Gerhard, O. E. Two-integral models of oblate elliptical galaxies with cusps. *Mon. Not. R. Astron. Soc.* **268**, 1019–1032 (1994).
58. Becker, R. H., White, R. L. & Helfand, D. J. The FIRST survey: faint images of the radio sky at twenty centimeters. *Astrophys. J.* **450**, 559–577 (1995).
59. Kennicutt, R. C. & Evans, N. J. Star formation in the Milky Way and nearby galaxies. *Annu. Rev. Astron. Astrophys.* **50**, 531–608 (2012).
60. Best, P. N. & Heckman, T. M. On the fundamental dichotomy in the local radio-AGN population: accretion, evolution and host galaxy properties. *Mon. Not. R. Astron. Soc.* **421**, 1569–1582 (2012).
61. Heckman, T. M. *et al.* Present-day growth of black holes and bulges: the Sloan Digital Sky Survey perspective. *Astrophys. J.* **613**, 109–118 (2004).
62. Cavagnolo, K. W. *et al.* A relationship between AGN jet power and radio power. *Astrophys. J.* **720**, 1066–1072 (2010).
63. McConnell, N. J. & Ma, C.-P. Revisiting the scaling relations of black hole masses and host galaxy properties. *Astrophys. J.* **764**, 184 (2013).
64. Osterbrock, D. E. *Astrophysics of Gaseous Nebulae and Active Galactic Nuclei* (University Science Books, Mill Valley, 1989).
65. Genzel, R. *et al.* The SINS survey of $z \sim 2$ galaxy kinematics: properties of the giant star-forming clumps. *Astrophys. J.* **733**, 101 (2011).
66. Carniani, S. *et al.* Ionised outflows in $z \sim 2.4$ quasar host galaxies. *Astron. Astrophys.* **580**, A102 (2015).
67. Sutherland, R. S. & Dopita, M. A. Cooling functions for low-density astrophysical plasmas. *Astrophys. J. Suppl. Ser.* **88**, 253–327 (1993).
68. Bohlin, R. C., Savage, B. D. & Drake, J. F. A survey of interstellar H I from L-alpha absorption measurements. II. *Astrophys. J.* **224**, 132–142 (1978).
69. Kennicutt, R. C. Jr *et al.* Star formation in NGC 5194 (M51a). II. The spatially resolved star formation law. *Astrophys. J.* **671**, 333–348 (2007).
70. Bershady, M. A. *et al.* The DiskMass survey. I. Overview. *Astrophys. J.* **716**, 198–233 (2010).

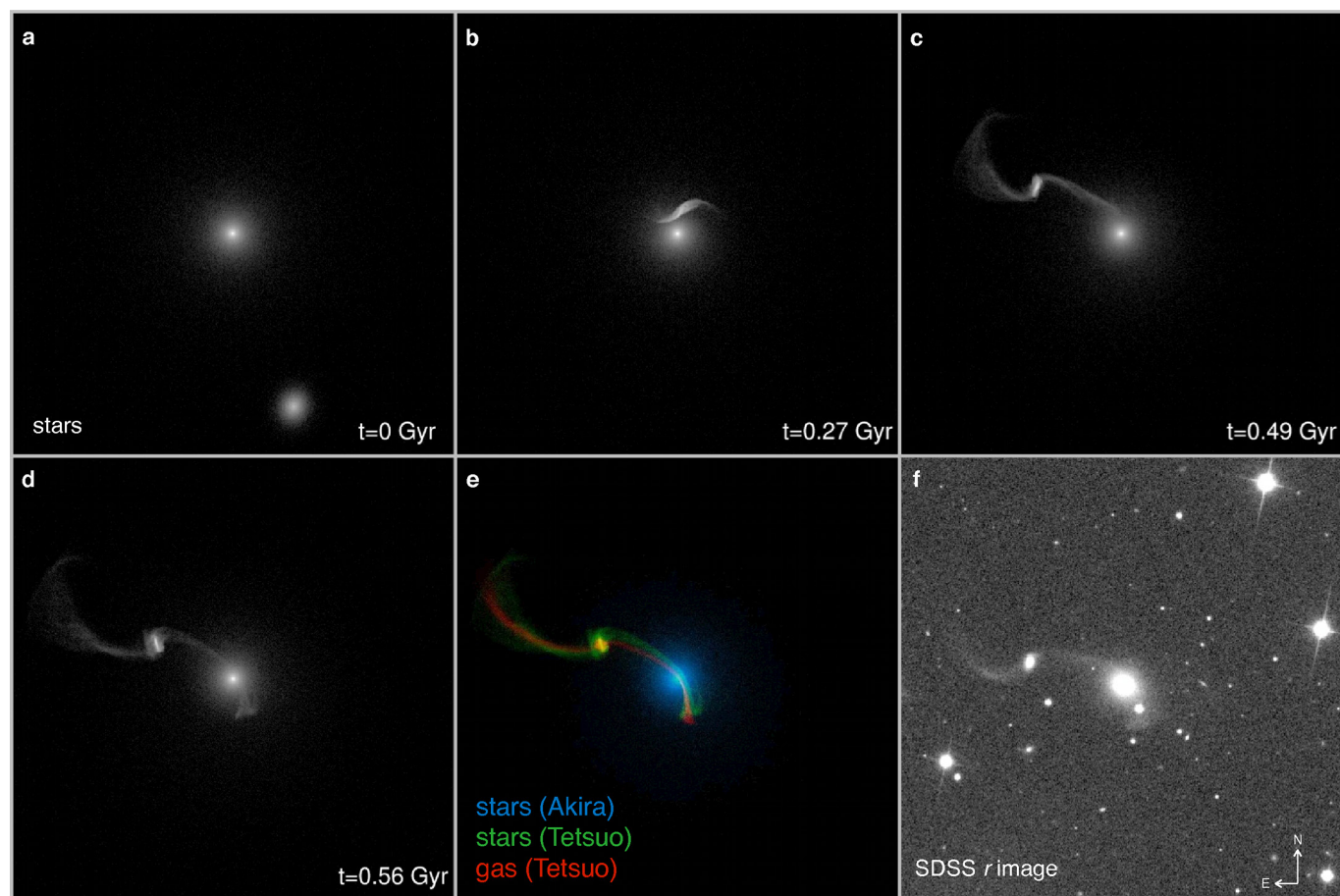


Extended Data Figure 1 | $A(V)$ map. The estimated $A(V)$ map (with $A(V)$ colour coded, see key), with contours of Na D EW $> 3.5 \text{ \AA}$ from Fig. 1d. The spatial overlap between regions of high extinction and the Na D EW absorption confirms that there is cool material in the foreground of Akira. Here and below, axes show offset in arcsec from map centre, marked with a cross.

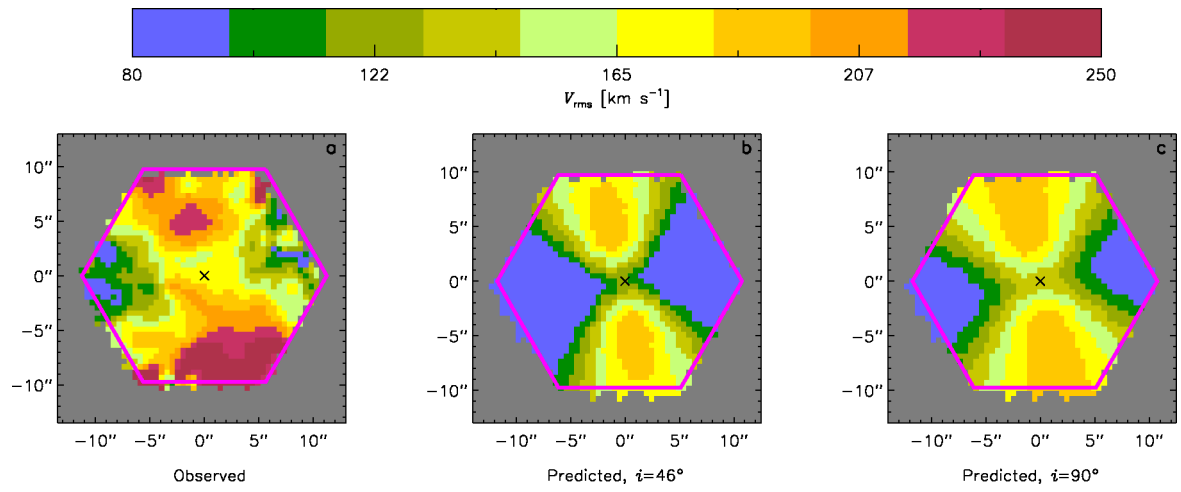


Extended Data Figure 2 | Na D line-of-sight measurement. **a–c**, The spectrum around the Na D doublet at $\lambda = 5,890, 5,896$ Å and best-fit stellar continuum. The two vertical lines mark the locations of the Na D doublet. **d–f**, The residual of the spectrum and stellar continuum. Considering only the wavelength range enclosed by the green region, we calculate the residual-weighted central wavelengths of these Na D doublets, which is

marked by the dashed grey vertical line and blue vertical lines. The dashed grey vertical represents the reference Na D centroid while the blue vertical lines represent the observed Na D centroid from the two spaxels of Akira. See Methods for details. The horizontal dashed line is a reference point and the Δv in **e** and **f** represents the residual-weighted velocities. Data in **a** from ref. 43 with permission.



Extended Data Figure 3 | Merger simulation. **a–d**, Evolution of the stars from $t=0$ Gyr to $t=0.56$ Gyr; each panel is 90×90 kpc. **e**, Composite image of stars and gas at $t=0.56$ Gyr; this panel is also 90×90 kpc. **f**, The SDSS *r* image of Akira and Tetsuo.



Extended Data Figure 4 | V_{rms} maps. **a**, Observed V_{rms} map (key at top shows colour coded V_{rms}). **b**, Predicted V_{rms} map, assuming $i = 46^\circ$. **c**, Predicted V_{rms} map, assuming $i = 90^\circ$.

A resonant chain of four transiting, sub-Neptune planets

Sean M. Mills¹, Daniel C. Fabrycky¹, Cezary Migaszewski^{2,3}, Eric B. Ford^{4,5,6}, Erik Petigura^{7,8} & Howard Isaacson⁷

Surveys have revealed many multi-planet systems containing super-Earths and Neptunes in orbits of a few days to a few months¹. There is debate whether *in situ* assembly² or inward migration is the dominant mechanism of the formation of such planetary systems. Simulations suggest that migration creates tightly packed systems with planets whose orbital periods may be expressed as ratios of small integers (resonances)^{3–5}, often in a many-planet series (chain)⁶. In the hundreds of multi-planet systems of sub-Neptunes, more planet pairs are observed near resonances than would generally be expected⁷, but no individual system has hitherto been identified that must have been formed by migration. Proximity to resonance enables the detection of planets perturbing each other⁸. Here we report transit timing variations of the four planets in the Kepler-223 system, model these variations as resonant-angle librations, and compute the long-term stability of the resonant chain. The architecture of Kepler-223 is too finely tuned to have been formed by scattering, and our numerical simulations demonstrate that its properties are natural outcomes of the migration hypothesis. Similar systems could be destabilized by any of several mechanisms^{5,9–11}, contributing to the observed orbital-period distribution, where many planets are not in resonances. Planetesimal interactions in particular are thought to be responsible for establishing the current orbits of the four giant planets in the Solar System by disrupting a theoretical initial resonant chain¹² similar to that observed in Kepler-223.

Kepler-223 is a known four-planet system¹³ orbiting around a slightly evolved (about 6-Gyr-old), Sun-like star (see Methods, Extended Data Fig. 1). The low observational signal-to-noise ratio initially caused an incorrect identification of the orbital periods of this system^{13,14}, and has hitherto precluded its detailed characterization. For the analysis of transit timing variation (TTV), we use long cadence (29.4-min

integrations) data, collected over the full duration of NASA's Kepler Space Mission from March 2009 to May 2013. Over this window, the ratios of the orbital periods (P) of planets b, c, d and e (named in alphabetic order from the interior, beginning with b) average $P_c/P_b = 1.3336$, $P_d/P_c = 1.5015$ and $P_e/P_d = 1.3339$ (ref. 15). We expect a system with periods so close to resonance to exhibit TTVs due to planet–planet interactions⁸ (see Methods).

To measure TTVs, we bin the data into 3-month segments based on Kepler's observing quarters, confirm that the orbital periods are near resonances, and demonstrate the time-variable nature of the transits (Fig. 1, Extended Data Fig. 2, Extended Data Table 1 and Methods). Phase folding the data and removing the TTVs allows the noisy transits to be identified easily by eye (Fig. 2).

The behaviour of the resonant chain can be characterized by its Laplace angles: $\phi_1 \equiv -\lambda_b + 2\lambda_c - \lambda_d$, $\phi_2 \equiv \lambda_c - 3\lambda_d + 2\lambda_e$ (for mean longitudes λ_i and planets $i = b, c, d, e$) and, for the whole system of four planets, $\phi_3 \equiv 2\phi_2 - 3\phi_1 = 3\lambda_b - 4\lambda_c - 3\lambda_d + 4\lambda_e$. Systems that are in resonance possess such librating Laplace angles, which ensures that two planets have a close approach when the other planets are far away, reducing chaotic interactions. The existence of a single four-body Laplace angle demonstrates that all the planets have close dynamical contact (with various three- and two-body resonances also present). We infer variations in the Laplace angles directly from the measured TTVs (see Methods and Extended Data Fig. 3). If we assume nearly circular orbits, the four years of TTVs in the data have recorded both angles performing nearly a full oscillation; ϕ_1 librates between approximately 173° and 190° and ϕ_2 librates between approximately 47° and 75°.

To improve the treatment of the TTV signal and directly connect it to planetary dynamics, we integrate the N -body equations of motion for the four-planet system and explicitly model the photometric transit signals over the Kepler observing window (photodynamical modelling)¹⁶.

Table 1 | Kepler-223 system parameters

Parameter name	DECMC result			
Spectroscopic stellar mass, M_* (M_\odot)	$1.125^{+0.094}_{-0.073}$			
Stellar radius, R_* (R_\odot)	$1.72^{+0.07}_{-0.14}$			
	Kepler-223 b	Kepler-223 c	Kepler-223 d	Kepler-223 e
Orbital period, P (d)	$7.38449^{+0.00022}_{-0.00022}$	$9.84564^{+0.00052}_{-0.00051}$	$14.78869^{+0.00030}_{-0.00027}$	$19.72567^{+0.00055}_{-0.00054}$
Eccentricity, e	$0.078^{+0.015}_{-0.017}$	$0.150^{+0.019}_{-0.051}$	$0.037^{+0.018}_{-0.017}$	$0.051^{+0.019}_{-0.019}$
Inclination, $ i - 90 $ (°)	$0.0^{+1.8}$	$0.0^{+1.3}$	$2.06^{+0.26}_{-0.32}$	$2.00^{+0.21}_{-0.21}$
Mass, M (M_\oplus)	$7.4^{+1.3}_{-1.1}$	$5.1^{+1.7}_{-1.1}$	$8.0^{+1.5}_{-1.3}$	$4.8^{+1.4}_{-1.2}$
Radius, R (R_\oplus)	$2.99^{+0.18}_{-0.27}$	$3.44^{+0.20}_{-0.30}$	$5.24^{+0.26}_{-0.45}$	$4.60^{+0.27}_{-0.41}$
Density, ρ (g cm^{-3})	$1.54^{+0.63}_{-0.35}$	$0.71^{+0.33}_{-0.20}$	$0.31^{+0.12}_{-0.07}$	$0.28^{+0.12}_{-0.08}$

Medians and 68% credible intervals for planet properties based on 2,008 10⁶-year stable solutions with eccentricity priors as described in Methods: ($e_{b,\text{max}}, e_{c,\text{max}}, e_{d,\text{max}}, e_{e,\text{max}}$) = (0.212, 0.175, 0.212, 0.175) and fixed nodal angle $\Omega_j = 0$ for $j = b, c, d, e$. All values are valid at an epoch time $T_{\text{epoch}} = 800.0$ (BJD – 2,454,900). The stellar mass (M_*) was held fixed in the differential-evolution Markov chain Monte Carlo (DECMC) simulation, but uncertainties in planetary mass were adjusted afterward to account for the quoted spectroscopic uncertainty in M_* . M_\odot and R_\odot are the mass and radius of the Sun, respectively; M_\oplus and R_\oplus are the mass and radius of Earth. See Methods and Extended Data Table 2 for additional parameters and discussion.

¹Department of Astronomy and Astrophysics, The University of Chicago, 5640 South Ellis Avenue, Chicago, Illinois 60637, USA. ²Institute of Physics and CASA*, University of Szczecin, Wielkopolska 15, 70-451 Szczecin, Poland. ³Torun Centre for Astronomy, Nicolaus Copernicus University, Gagarina 11, 87-100 Torun, Poland. ⁴Center for Exoplanets and Habitable Worlds, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. ⁵Department of Astronomy and Astrophysics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. ⁶Center for Astrostatistics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. ⁷University of California at Berkeley, Berkeley, California 94720, USA. ⁸California Institute of Technology, Pasadena, California 91125, USA.

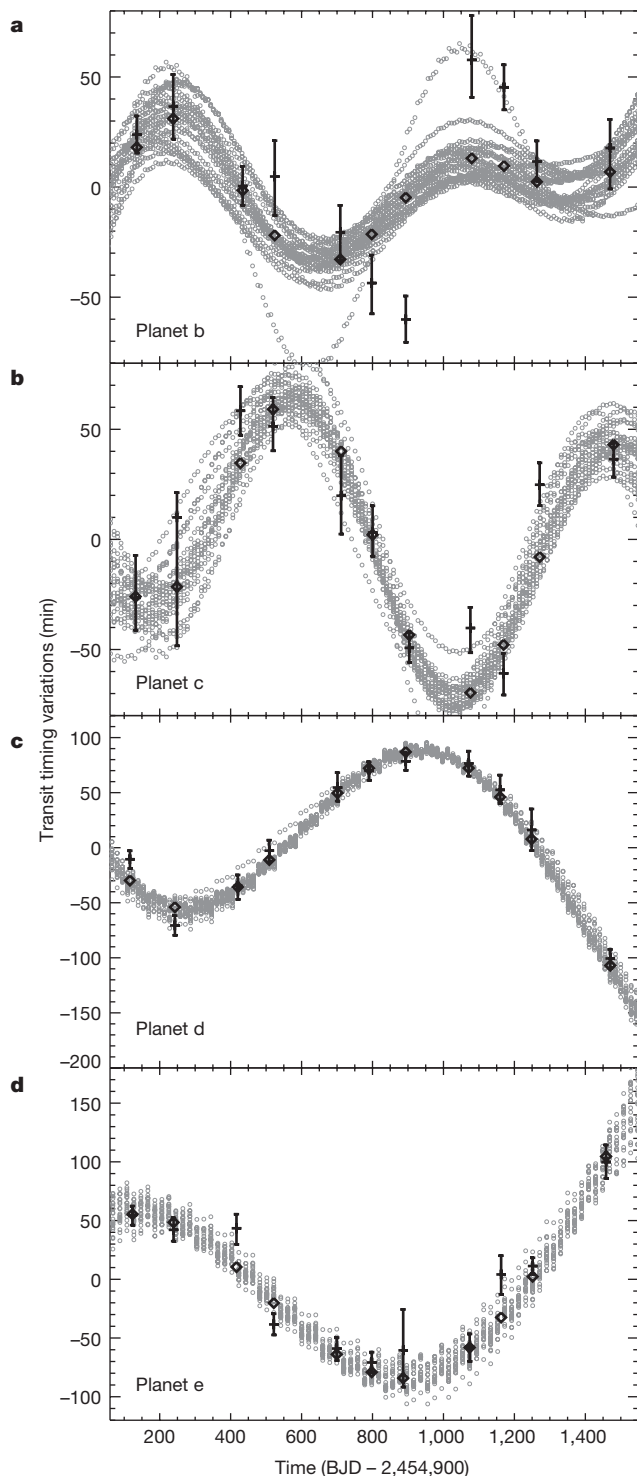


Figure 1 | Transit timing variations (TTVs) for all four planets with respect to a linear ephemeris. a–d, Calculated transit times for planets b–e, respectively, come from a linear regression of the best-fit model transits. Open grey circles show the transit times from 20 different models that were stable over a 10^7 -year simulation. Black '+' symbols with 1σ error bars indicate the TTVs found by fitting quarterly binned data (see Extended Data Fig. 2), and black diamonds are the corresponding points for the mean of the grey-circle models binned in the same manner. Where the noise causes large uncertainties, the photodynamic model may deviate from the binned data, but more accurately reflects the true TTVs. BJD, barycentric Julian date.

We determine best estimates and uncertainties for the system parameters by performing five-body integrations of initial conditions from the resulting posterior distribution for more than 10^7 orbits of the planets

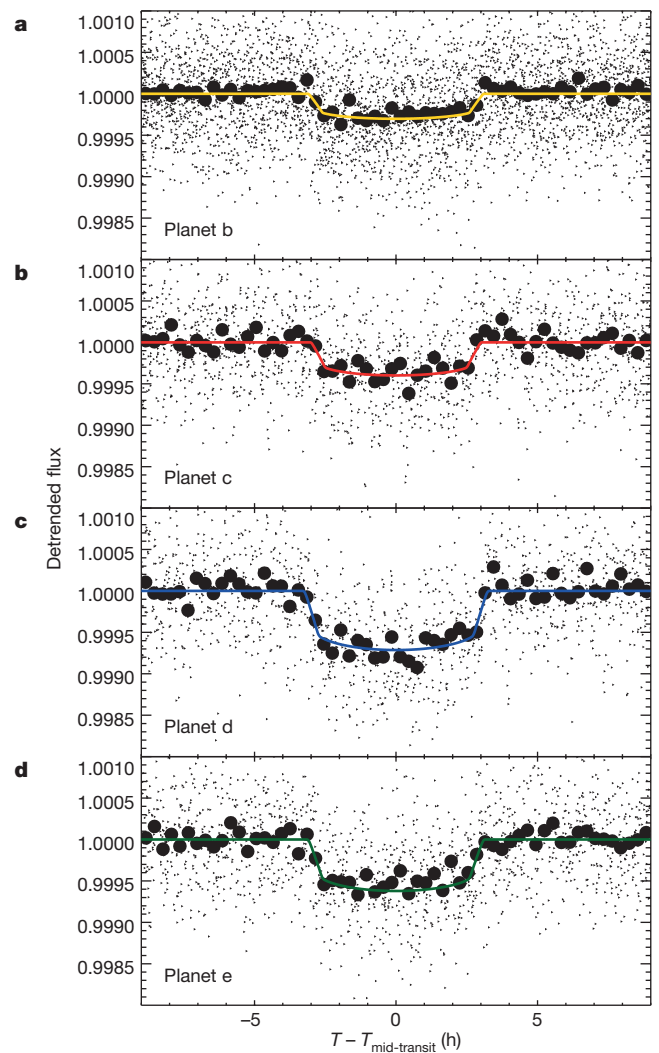


Figure 2 | Binned planet transits. a–d, Photometry data near transits of planets b–e, respectively (small black triangles), binned together (large black circles) by phase-folding after removing the measured TTV for each quarter. Systematic trends have been removed and the flux normalized to 1.0 out of transit. The coloured lines are the best-fit transit models to the data.

and retaining only parameter sets that remain stable (see Methods for details). We find that the planets all have masses of $3M_{\oplus}$ – $9M_{\oplus}$ and radii of $2.5R_{\oplus}$ – $5.5R_{\oplus}$ (M_{\oplus} and R_{\oplus} are the mass and radius of Earth, respectively; see Table 1). On the basis of these values and internal structure models¹⁷, we determine that the composition of the planets varies from about 1% to 5% H/He by mass for the innermost planet to more than about 10% by mass for the outermost planet; that is, they are all sub-Neptunes. The density of the planets decreases with orbital semi-major axis, consistent with scenarios involving atmospheric loss due to stellar irradiation or formation in regions of increasingly cooler temperatures¹⁸. The eccentricities of the planets are relatively low (about 0.01–0.1) in configurations that are stable for more than 10^7 orbits of the system. To fit the data acceptably, the eccentricities need to be slightly larger than in other systems of sub-Neptunes such as Kepler-11, whose eccentricities are less than about 0.02 (ref. 19). Because the eccentricities may be excited and stabilized by the resonances, the system can remain stable even though it is compact. The eccentricity of a planet is only loosely negatively correlated with its mass (from the TTVs in the data), so small changes in the allowed eccentricity will have a small effect on the posterior mass estimate, and removing eccentricity constraints would make the planets only slightly less dense.

Periods in a ratio close to 3:4:6:8 are maintained in all the stable, data-fitting solutions. The range of the ratios of the osculating periods

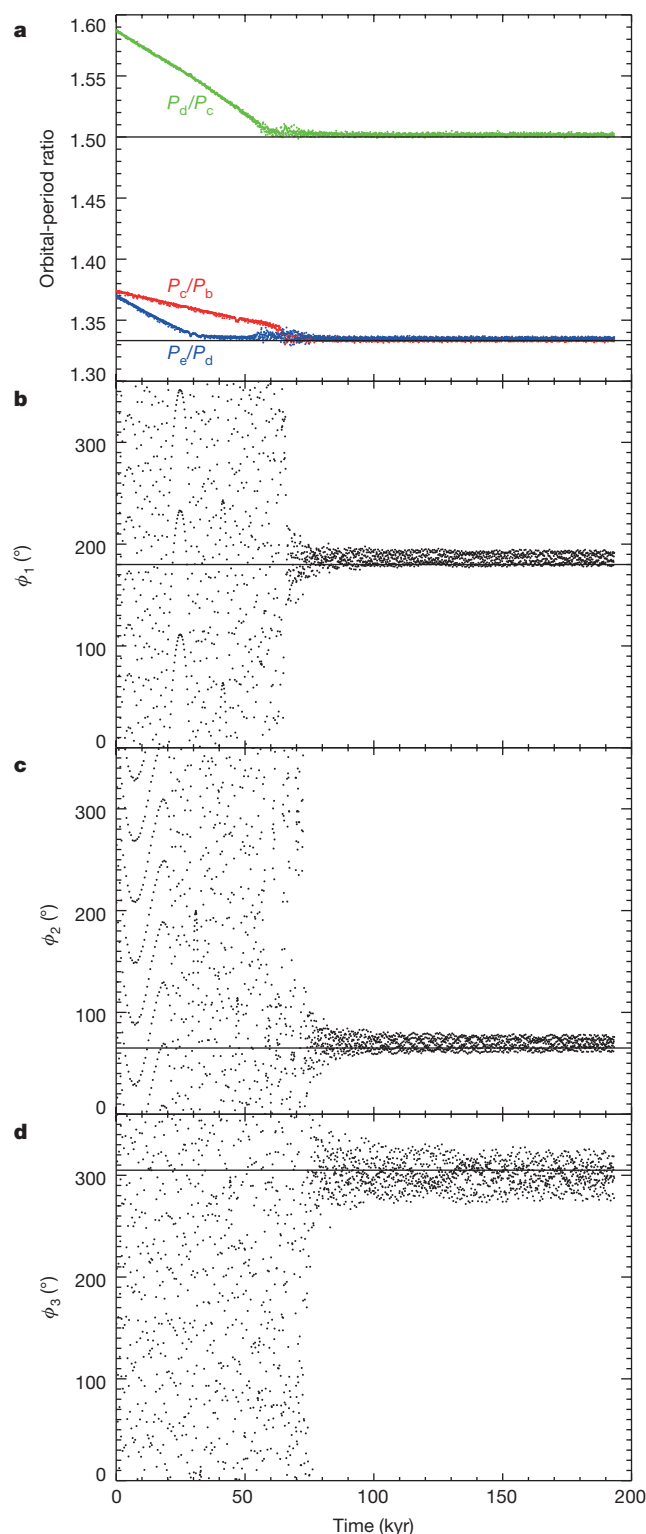


Figure 3 | A migration simulation that ends in a configuration matching the observed semi-major axis ratios, and libration angle centres and amplitudes. **a**, Time evolution of orbital-period ratios of planets b and c (P_c/P_b ; red), c and d (P_d/P_c ; green), and d and e (P_e/P_d ; blue) in a migration simulation. **b–d**, Time evolution of the Laplace angles (ϕ_{1-3}) defined in the text. The resonant angles and libration amplitudes that the planets end up in (indicated by the black horizontal lines) match those observed in the data (see, for example, Extended Data Fig. 3).

of the planets implied by the observed TTVs over the Kepler window is typical for a resonant system. This range is narrower than that for a long-lived (more than about 10^7 orbits), but circulating

(non-resonant), solution (Extended Data Figs 4, 5), suggesting that the system is currently in a state of libration. This libration might be temporary, and periods of Laplace-angle circulation might have occurred previously or might occur in the future for this system. However, requiring short-term Laplace-angle libration substantially increases the likelihood that a parameter set that acceptably fits the data represents a long-lived system (see Methods). Because (i) the orbital parameters of Kepler-223 are consistent with it being in a resonant state, (ii) solutions that are stable for 100 Myr exist within the parameter posteriors, and (iii) resonance greatly helps a system this compact to remain stable, we conclude that the system is probably a true resonant chain.

Planetary migration in a disk has been extensively studied and often leads to resonant chains of planets^{3–6}. To examine the plausibility of the specific resonant chain observed in Kepler-223, we use a previously developed model²⁰ to simulate the migration of four planets within a gas disk. We find that four planets starting well wide of resonance migrate inwards and converge to the 3:4:6:8 chain of periods that we observe with certain choices of simulation parameters (Fig. 3). Thus the Kepler-223 system is a plausible outcome of disk migration, but the full set of disk migration parameters and initial conditions that would lead to this system remains an open question.

In a migration scenario, systems trapped in resonances for which the orbital semi-major axes are small (less than about 0.5 AU) can potentially be used to constrain the rate of disk photoevaporation and the lifetimes of disks, because a gaseous disk must exist in the 0.02–0.2 AU range long enough for planets of moderate mass to migrate. It also provides constraints on turbulence and magnetic fields in the disk²¹, and the structure of the disk that causes the planets to stop migrating²². An alternative to gas-disk migration for trapping planets into resonances is migration via planetesimal scattering²³. It is possible for planetesimal scattering to migrate two planets in a convergent manner, establishing a resonance. However, this convergent migration would excite the eccentricities of the planetesimal population, which would probably prevent additional planets from joining the resonance²⁴. The presence of a large volatile (greater than about 10% H/He by mass)¹⁷ layer on the outer planets also suggests that the planets formed in the presence of a gas-containing disk at cool temperatures, further suggesting large-scale migration¹⁸.

Several other exoplanet systems have (GJ 876; ref. 25), or are speculated to have (HR 8799; ref. 20), resonant chains, but these are composed of planets that are substantially more massive and have much greater orbital distances; hence, these observations may not be relevant to the formation of systems of close-in sub-Neptunes. Several Kepler systems are probably in a true resonance (as opposed to near resonance; for example, the 6:5 system Kepler-50 and the 5:4:3 system Kepler-60; ref. 26); however, owing to the large number of known multi-planet systems, even if the orbital-period ratios of planets are essentially random, consistent with *in situ*, giant-impact formation, we would expect to observe some systems whose period ratios were near enough to integer values that they entered true dynamical resonances. By contrast, the precise conditions for the four-planet resonant chain of Kepler-223 cannot be accounted for by random selection of period ratios⁷, and the system is probably too fragile to have been assembled by giant impacts²⁷.

The dynamical fragility of Kepler-223 suggests that resonant chains were precursors to some of the more common, non-resonant systems and that planet–planet scattering post-formation is probably an important step in creating the observed period distribution¹⁰. A model of the formation of the Solar System that has parallels with observed exoplanets involves the four giant planets entering a series of resonances, reaching their current configuration only after destabilization hundreds of millions of years later¹². Numerical simulations for Kepler-223 indicate that only a small mass of orbit-crossing planetesimals is needed to move Kepler-223 off resonance²⁸, but that it could escape this fate if intrinsic differences in protoplanetary disks resulted in the lack of such a planetesimal population. In fact, various mechanisms including disk

dissipation⁹, planet–planet scattering¹⁰, tidal dissipation⁵ and planetesimal scattering¹¹ could break migration-induced resonances in the majority of exoplanet systems. It has been suggested that some multi-resonant systems (for example, Kepler-80, which has planetary pairs near, but not in, two-body resonances) might have undergone resonant disruption as a result of tidal dissipation, which would explain most of the period ratios that are slightly greater than resonant values in Kepler data^{29,30}. It is possible that the Kepler-223 resonance has survived as a result of its relatively more distant innermost planet. Overall, we suggest that substantial migration of planets, including epochs of resonance that are typically only temporary, rather than *in situ* formation, leads to the final, observed planetary orbits for many close-in sub-Neptune systems.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 August 2015; accepted 11 February 2016.

Published online 11 May 2016.

- Mullally, F. *et al.* Planetary candidates observed by *Kepler*. VI. Planet sample from Q1–Q16 (47 months). *Astrophys. J. Suppl. Ser.* **217**, 31 (2015).
- Hansen, B. M. S. & Murray, N. Testing *in situ* assembly with the *Kepler* planet candidate sample. *Astrophys. J.* **775**, 53 (2013).
- Melita, M. D. & Woolfson, M. M. Planetary commensurabilities driven by accretion and dynamical friction. *Mon. Not. R. Astron. Soc.* **280**, 854–862 (1996).
- Lee, M. H. & Peale, S. J. Dynamics and origin of the 2:1 orbital resonances of the GJ 876 planets. *Astrophys. J.* **567**, 596–609 (2002).
- Terquem, C. & Papaloizou, J. C. B. Migration and the formation of systems of hot super-Earths and Neptunes. *Astrophys. J.* **654**, 1110–1120 (2007).
- Cresswell, P. & Nelson, R. P. On the evolution of multiple protoplanets embedded in a protostellar disc. *Astron. Astrophys.* **450**, 833–853 (2006).
- Fabrycky, D. C. *et al.* Architecture of *Kepler*'s multi-transiting systems. II. New investigations with twice as many candidates. *Astrophys. J.* **790**, 146 (2014).
- Agol, E., Steffen, J., Sari, R. & Clarkson, W. On detecting terrestrial planets with timing of giant planet transits. *Mon. Not. R. Astron. Soc.* **359**, 567–579 (2005).
- Cossou, C., Raymond, S. N., Hersant, F. & Pierens, A. Hot super-Earths and giant planet cores from different migration histories. *Astron. Astrophys.* **569**, A56 (2014).
- Pu, B. & Wu, Y. Spacing of *Kepler* planets: sculpting by dynamical instability. *Astrophys. J.* **807**, 44 (2015).
- Chatterjee, S. & Ford, E. B. Planetesimal interactions can explain the mysterious period ratios of small near-resonant planets. *Astrophys. J.* **803**, 33 (2015).
- Levison, H. F., Morbidelli, A., Tsiganis, K., Nesvorný, D. & Gomes, R. Late orbital instabilities in the outer planets induced by interaction with a self-gravitating planetesimal disk. *Astron. J.* **142**, 152 (2011).
- Borucki, W. J. *et al.* Characteristics of planetary candidates observed by *Kepler*. II. Analysis of the first four months of data. *Astrophys. J.* **736**, 19 (2011).
- Lissauer, J. J. *et al.* Architecture and dynamics of *Kepler*'s candidate multiple transiting planet systems. *Astrophys. J. Suppl. Ser.* **197**, 8 (2011).
- Lissauer, J. J. *et al.* Validation of *Kepler*'s multiple planet candidates. II. Refined statistical framework and descriptions of systems of special interest. *Astrophys. J.* **784**, 44 (2014).
- Carter, J. A. *et al.* Kepler-36: a pair of planets with neighboring orbits and dissimilar densities. *Science* **337**, 556–559 (2012).
- Lopez, E. D. & Fortney, J. J. Understanding the mass-radius relation for sub-Neptunes: radius as a proxy for composition. *Astrophys. J.* **792**, 1 (2014).
- Lee, E. J. & Chiang, E. Breeding super-Earths and birthing super-puffs in transitional disks. *Astrophys. J.* **817**, 90 (2016).
- Lissauer, J. J. *et al.* A closely packed system of low-mass, low-density planets transiting Kepler-11. *Nature* **470**, 53–58 (2011).
- Goździewski, K. & Migaszewski, C. Multiple mean motion resonances in the HR 8799 planetary system. *Mon. Not. R. Astron. Soc.* **440**, 3140–3171 (2014).
- Adams, F. C., Laughlin, G. & Bloch, A. M. Turbulence implies that mean motion resonances are rare. *Astrophys. J.* **683**, 1117–1128 (2008).
- Masset, F. S., Morbidelli, A., Crida, A. & Ferreira, J. Disk surface density transitions as protoplanet traps. *Astrophys. J.* **642**, 478–487 (2006).
- Minton, D. A. & Levison, H. F. Planetesimal-driven migration of terrestrial planet embryos. *Icarus* **232**, 118–132 (2014).
- Ormel, C. W., Ida, S. & Tanaka, H. Migration rates of planets due to scattering of planetesimals. *Astrophys. J.* **758**, 80 (2012).
- Nelson, B. E. *et al.* An empirically derived three-dimensional Laplace resonance in the Gliese 876 planetary system. *Mon. Not. R. Astron. Soc.* **455**, 2484–2499 (2016).
- Goździewski, K., Migaszewski, C., Panichi, F. & Szuszkiewicz, E. The Laplace resonance in the Kepler-60 planetary system. *Mon. Not. R. Astron. Soc.* **455**, L104–L108 (2016).
- Raymond, S. N., Barnes, R., Armitage, P. J. & Gorelick, N. Mean motion resonances from planet-planet scattering. *Astrophys. J.* **687**, L107–L110 (2008).
- Moore, A., Hasan, I. & Quillen, A. C. Limits on orbit-crossing planetesimals in the resonant multiple planet system, KOI-730. *Mon. Not. R. Astron. Soc.* **432**, 1196–1202 (2013).
- Batygin, K. & Morbidelli, A. Dissipative divergence of resonant orbits. *Astron. J.* **145**, 1 (2013).
- Delisle, J.-B. & Laskar, J. Tidal dissipation and the formation of *Kepler* near-resonant planets. *Astron. Astrophys.* **570**, L7 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank A. Howard and G. Marcy for their role in obtaining spectra, and E. Agol, J. Lissauer, and J. Bean for comments on the manuscript. This material is based on work supported by NASA under grant numbers NNX14AB87G (D.C.F.), NNX12AF73G (E.B.F.) and NNX14AN76G (E.B.F.) issued through the Kepler Participating Scientist Program. E.B.F. received support from NASA Exoplanet Research Program award NNX15AE21G. D.C.F. received support from the Alfred P. Sloan Foundation. C.M. was supported by the Polish National Science Centre MAESTRO grant DEC-2012/06/A/ST9/00276.

Author Contributions S.M.M. performed the photodynamic, stability, tidal dissipation and spectral evolution analyses and led the paper authorship. D.C.F. designed the study, performed TTV and Laplace-angle libration analysis, and assisted writing the paper. C.M. performed the migration analysis, assisted in initial data fitting and contributed to the writing of the paper. E.B.F. advised on the DEMCMC analysis and paper direction. E.P. and H.I. obtained and analysed the spectra. All authors read and edited the manuscript.

Author Information Kepler data are publicly available at <http://archive.stsci.edu/kepler/>. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.M.M. (sean.martin.mills@gmail.com).

METHODS

Stellar properties. To improve our knowledge of the Kepler-223 system, we obtained a spectrum of the host star on 10 April 2012 using the HIRES spectrometer³¹ at the Keck-1 10-m telescope. These data are now publicly available at <http://cfop.ipac.caltech.edu>. After normalizing the continuum, we model the observed spectrum using synthetic spectra. Model spectra are generated by interpolating within a grid of synthetic spectra³². The resulting spectroscopic parameters for Kepler-223 are $T_{\text{eff}} = 5,821 \pm 123$ K, $\log(g) = 4.070 \pm 0.096$ dex and $[\text{Fe}/\text{H}] = 0.060 \pm 0.047$ dex (where g is the surface gravity in cm s^{-2} and the metallicity $[\text{Fe}/\text{H}]$ is the logarithm of the ratio of iron to hydrogen in the star relative to that ratio in the sun).

To determine an age and mass of the star, we match the measured properties to Y^2 isochrones³³. We ran a Markov chain Monte Carlo (MCMC) using the spectroscopic data and an interpolation of the Y^2 grid values as the model to obtain an age of $6.3^{+1.8}_{-1.7}$ Gyr and mass of $1.125^{+0.095}_{-0.073} M_{\odot}$ (see Extended Data Fig. 1). Combining these values with $\log(g)$, we measure the stellar radius, $R_{\star} = 1.54^{+0.21}_{-0.18} R_{\odot}$, and stellar density, $\rho_{\star} = 0.31^{+0.12}_{-0.09} \rho_{\odot}$. We also derive a distance from Earth of $2.29^{+0.34}_{-0.34}$ kpc to Kepler-223 and find the mean flux S on the planets to be $S_b = (492 \pm 47) S_0$, $S_c = (335 \pm 32) S_0$, $S_d = (195 \pm 19) S_0$ and $S_e = (133 \pm 13) S_0$, where $S_0 = 1,377 \text{ W m}^{-2}$ is the average insolation of Earth.

To determine the size of model-dependent uncertainties, we compare our results to an independently developed, publicly available method for computing M_{\star} , R_{\star} and age using the Dartmouth isochrones³⁴ (<https://github.com/timothydmorton/isochrones>). All three values are consistent within the 1σ error bars, so we conclude that our measurements are robust and that model-dependent errors are small compared to our quoted uncertainties. We also use a stellar population synthesis model, TRILEGAL³⁵, with the default galaxy stellar distribution and population as described therein, to demonstrate that the best-fit mass and uncertainties described above are essentially unaffected by reasonable priors; so, we keep flat priors for all stellar parameters.

TTVs. To measure TTVs, we begin by detrending the simple aperture photometry (SAP) flux data from the Kepler portal on the Mikulski Archive for Space Telescopes (MAST). For long-cadence data (quarters 1–8), we fit the amplitudes of the first five co-trending basis vectors (largest magnitude vectors from a singular value decomposition of the photometry for a given CCD channel) to determine a baseline. We discard points marked as low quality (quality flag of ≥ 16). For short-cadence data (58.8-s integrations, quarters 9–17), co-trending basis vectors are not available. Instead, we first masked out the expected transit times of a preliminary model, plus 20% of the full duration of each transit intending to account for possible additional timing variations; then we fit a cubic polynomial model with a 2-day width centred within half an hour of each data point to determine its baseline. In both cases, the baseline remains dominated by instrumental systematics that are time-variable; thus, we divide the flux by this baseline.

In computing TTVs, we use only those data for which the transits do not overlap with another planetary transit (that is, two transit mid-times fall within 1 day of each other) according to a preliminary model (data with overlapping transits is modelled directly by the photodynamic method described later). To determine transit times, we first fit transit parameters (period, transit mid-time, planet-to-star radius ratio, transit duration, impact parameter and limb-darkening coefficient) to the entire long-cadence dataset. Second, we refit each quarter using the globally determined values for all parameters except for transit mid-time, which is solved for. Third, we refine the transit shape parameters and slide the refined transit model in time through the data for each planet in each quarter, computing the goodness-of-fit statistic χ^2 in steps of 0.001 days. The values of the numerical χ^2 function that are within 1.0 of the minimum are fit with a parabola, the minimum of which we adopt as our best estimate of the mid-time. The time shifts in each direction at which the χ^2 function rises by 1 and 9 above the minimum are adopted as narrow and conservative error bars. If the likelihood surface of the mid-time parameter was Gaussian, these values would correspond to 1σ and 3σ estimates. Extended Data Table 1 reports the average time of the transits that were combined to make each measurement, the best-estimate and uncertainty estimates of these time shifts. Once phased at these transit times, the transit light curves are shown in Fig. 2. These transit times are also represented graphically in Extended Data Fig. 2 as the horizontal error bar. Planets c, d and e all have visible fluctuations over the dataset. These data constitute our transit timing measurement, which does not depend on the photodynamical model we develop subsequently; also, the data are not used in this model.

We use these transit times to estimate the Laplace critical angles³⁶ and their evolution. To do so, we note that for circular orbits the mean longitude, λ , is a linear function of time, t , related to the transit period, P , and a specific mid-time, T'_0 , as

$$\lambda = 2\pi[1/4 + (t - T'_0)/P]$$

In place of T'_0 we may use $T_0 + \Delta T_0$, where P and T_0 define the linear ephemeris on which the quarterly ΔT_0 of Extended Data Table 1 are based. Then, for Laplace's critical angles we have

$$\begin{aligned} \phi_1 &= -\lambda_b + 2\lambda_c - \lambda_d \\ &= 2\pi \left[\frac{T'_{0b}}{P_b} - \frac{2T'_{0c}}{P_c} + \frac{T'_{0d}}{P_d} + t \left(\frac{2}{P_c} - \frac{1}{P_b} - \frac{1}{P_d} \right) \right] \\ &= 2\pi \left[0.4750 + 2.39834 \times 10^{-5}(t - 2,454,900) + \frac{\Delta T_{0b}}{P_b} - \frac{2\Delta T_{0c}}{P_c} + \frac{\Delta T_{0d}}{P_d} \right] \end{aligned}$$

where t is given in units of days in terms of the barycentric Julian date (BJD), and similarly

$$\begin{aligned} \phi_2 &= \lambda_c - 3\lambda_d + 2\lambda_e \\ &= 2\pi \left[-\frac{T'_{0c}}{P_c} + \frac{3T'_{0d}}{P_d} - \frac{2T'_{0e}}{P_e} + t \left(\frac{1}{P_c} - \frac{3}{P_d} + \frac{2}{P_e} \right) \right] \\ &= 2\pi \left[0.1135 + 8.7366 \times 10^{-5}(t - 2,454,900) - \frac{\Delta T_{0c}}{P_c} + \frac{3\Delta T_{0d}}{P_d} - \frac{2\Delta T_{0e}}{P_e} \right] \end{aligned}$$

These values are plotted in Extended Data Fig 3. The values are not constant, and the data appear to have sampled a minimum and maximum value of a libration cycle, which indicates a restoring torque. The specific values are sensitive to phase shifts due to eccentricity-vector precession; the libration centres may be different by about 30° if the eccentricities are as high as 0.1.

Photodynamic inputs. A Newtonian photodynamic model similar to existing models³⁷, but developed independently, was used for a dynamical analysis of this system. To find the most likely parameter values and uncertainties in the system, we run a differential-evolution Markov chain Monte Carlo (DEMCMC)³⁸ to compare model output for different system parameters to observed long- and short-cadence Kepler data, as well as spectroscopic data of the star. The TTV signal (Fig. 1), which here is constrained by the photometry directly, detects the gravitational perturbations due to planet mass. Combined with transit shape information, this constrains the eccentricities and provides positive mass detections at $>2.5\sigma$ for all bodies with uncertainties approximately 10–30% of the fitted values.

Each planet ($i = b, c, d, e$) has seven parameters: $\mathbf{p}_i = [P, T_0, \text{ecos}(\omega), i, \Omega, R_p/R_{\star}, M_p/M_{\star}]$ in which P is the period, T_0 is the mid-transit time, e is the eccentricity, i is the inclination, ω is the argument of periastron, Ω is the nodal angle, R_p/R_{\star} is the planet-to-star radius ratio and M_p/M_{\star} is the planet-to-star mass ratio. The star has five parameters: $\mathbf{p}_{\star} = [M_{\star}, R_{\star}, c_1, c_2, \text{dilution}]$, in which c_i are the two quadratic limb-darkening coefficients and 'dilution' is the amount of dilution from other stars. Because photometry constrains only stellar density, and not mass and radius individually, we fix M_{\star} at the best-fit value found from spectroscopy and convolve the mass distribution with the DEMCMC posteriors when reporting final values.

We fix $\Omega = 0$ for all planets because the data do not sensitively measure mutual inclinations. The typical mean mutual inclination (MMI) of Kepler systems, approximately 1.8° , implies near coplanarity⁷. Additionally, multi-planet systems with higher mutual inclinations between planetary orbital planes are correlated with instability³⁹, and we expect any observed system to be at least quasi-stable. Although, for some pairs of planets, photometry determines whether their inclinations are on the same side of 90° (refs 40, 41), in preliminary runs we find no preference for either conclusion. Therefore, we explore only $i > 90^\circ$ for each planet to reduce the volume of the symmetric parameter space. The value for the stellar limb-darkening coefficient c_2 was chosen as 0.2 because this value is close to the median value for stars in the 4,000–6,500-K range in the Kepler bandpass⁴², and for low signal-to-noise ratio transits such as that in Kepler-223, a single limb-darkening parameter is sufficient to match transit shape^{43,44}.

United Kingdom Infrared Telescope (UKIRT) archives reveal that there are two objects within $2''$ of the position specified by the Kepler Input Catalog (KIC)⁴⁵. The brighter of the two objects is less than $0.2''$ from the KIC position and has a predicted Kepler magnitude of 15.4932, which is based on the formula used within the UKIRT archives to convert the measured J-band magnitudes to a Kepler magnitude⁴⁶. This value is 0.1492 magnitudes fainter than that reported in the KIC (15.344). The second object is $1.937''$ away from the KIC location, but is about 8 times fainter. The sum of these two objects has a predicted intensity in the Kepler bandpass equal to 98.2% of the intensity of the object reported by the KIC. Faulkes Telescope North (FTN) imaging confirms the dual nature of the Kepler-223 object⁴⁷. Speckle imaging done at WIYN observatory indicates no additional bodies between approximately $0.2''$ and $1.9''$ of the brighter object⁴⁸. Because the fainter of the two objects contributes approximately 11.202% of the light in the Kepler bandpass, we perform our DEMCMC runs with the dilution fixed at 0.11202.

Photodynamic fits. Beginning the DEMCMC by distributing parameters over the entire 30-dimensional prior is computationally untenable for this problem because it would take an excessively long time for the parameter sets, $\{p_i\}$, of the DEMCMC to escape local minima and reach the global minimum. Instead we begin the DEMCMC by taking a four-planet solution found by exploration using migration-assembly solutions, p_0 , which approximately matches the observed data, and forming a set of 48 30-parameter vectors, $\{p_0\}$, by adding 30-dimensional Gaussian noise to p_0 . We allow each set to explore the parameter space and, to eliminate any effects of the choice of p_0 , we wait until the DEMCMC chains have converged and then remove a ‘burn-in’ period, that is, the portion that is dependent on the choice of $\{p_0\}$.

In the DEMCMC, a given choice of planetary parameters is accepted or rejected one the basis of the data over the Kepler observing window (about 4 years), and does not take into account the long-term evolution of a system with such parameters. It is not computationally tenable to numerically integrate each model for the age of the Kepler-223 system during the DEMCMC run. Therefore, the DEMCMC posterior includes solutions that acceptably fit the data, but that become unstable shortly after. To prevent our posterior parameter estimates from representing unstable solutions, we take two steps to encourage stability. First, we do not allow the DEMCMC to explore any solutions where the orbits of two adjacent planets cross (which generates a posterior we call C_1). This was implemented by allowing the DEMCMC to explore a limited range of eccentricities for each planet, $(e_{b,\max}, e_{c,\max}, e_{d,\max}, e_{e,\max}) = (0.212, 0.175, 0.212, 0.175)$, with the symmetry of values due to the resonant-chain structure of the periods (posteriors can be found in Extended Data Table 2 and best-fits in Extended Data Table 3). Retrospectively, this eccentricity prior is justified because mean eccentricities greater than 0.1 are very rarely stable (Extended Data Fig. 6). Further, the similarity between the 10^6 -year eccentricity-stability distribution and the 10^7 -year distribution indicates that using either as a proxy for stable solutions will yield comparable results.

To assess the stability of the solutions in the posterior distribution, we selected 500 random draws from the C_1 posterior and numerically integrated each of these solutions for 10^7 years, which corresponds to more than 10^8 orbits of the outermost planet. We used the MERCURY symplectic integrator⁴⁹ and stopped integration if a close encounter between any two bodies occurred. 30% of systems lasted the entire 10^7 -year integration. We randomly selected 25 of the systems that lasted 10^7 years and numerically integrated them for an additional 9×10^7 years, or until a close encounter, with 64% of them lasting 10^8 years. The age of the Kepler-223 star is about 6×10^9 years. We expect the planets to have reached their current configuration by migration through a disk within only a few million years, which corresponds to the lifetimes of gas disks⁵⁰, suggesting that the current planet configuration has also survived for about 6×10^9 years. However, integrating for this long is not computationally feasible for this study. Other numerical stability studies¹⁰ predict that systems are approximately equally likely to become unstable in bins of $\log(\text{time})$, implying that approximately 12% of the tested systems (and thus approximately 12% of the systems in the C_1 posterior) remain stable on timescales of billions of years. This fraction is high compared to a modelled population of compact, sub-Neptune systems, which are destabilized by mean motion resonances (MMRs) on a shorter timescale¹⁰. However, in such simulations there are generally a few bodies not engaged in the resonance; here all four bodies are involved in the resonance, remaining stable despite MMRs exciting eccentricities. Also, MERCURY is a Newtonian physics integrator, but adding a suitable general relativistic potential term, $U_{\text{GR}} = -3(GM_*/(cr))^2$, where G is the gravitational constant, c is the speed of light and r is the distance from a planet to the star⁵¹, does not change our long-term stability results from 100 trials (32 stable, 68 unstable).

To develop a second posterior based on parameters that are more likely to lead to stability, we randomly drew 5,000 parameter sets from the posterior of C_1 , and numerically integrated each of these solutions for 10^6 years (corresponding to more than 10^7 orbits of the outermost planet). This allows the problem to be computationally feasible, while still allowing for a large enough number of draws that we have sufficient statistics for parameter estimates. We retained only those parameter sets that remained stable at least this long (2,008 in total) to form a second posterior representative of physical (stable) solutions and call it C_2 . Future discussions of parameters and the data in the main text (Table 1 and Fig. 1) use this posterior (C_2) because we judge it to be the optimal combination of selecting stable solutions that match the observed Kepler data, while avoiding discarding plausible parameter space as a result of further assumptions. The general shape of the eccentricity distribution remaining after 10^6 years does not change markedly compared to solutions that are stable for an order of magnitude longer (see Extended Data Fig. 6) and is thus unlikely to change noticeable over the ~ 6 -Gyr age of the system. The instability regions near the best-fit values discovered by our parameter fits suggest the ease with which the system, and others like it, could

be moved out of resonance by small perturbations such as evaporation of the protoplanetary disk^{9,28}.

Kepler-223 appears to possess two librating Laplace angles between the inner three and outer three planets, as discussed earlier. Migration simulations suggest that a very large Laplace-angle libration amplitude is unlikely in stable solutions. Further, in stable solutions in the C_2 posterior, long-lived (up to about 10^5 years) Laplace-angle libration is likely to occur. To get another estimate of the parameters of the system while balancing computational efficiency and a stricter stability constraint, we ran a third DEMCMC. For this run, at every step in the DEMCMC we integrate the parameter initial conditions for 100 years (corresponding to more than about 5 secular oscillations) and penalize Laplace-angle oscillation amplitudes that grow too large, in addition to fitting the data. We call the posterior from this run C_3 . Our Laplace-angle criteria in C_3 are designed to penalize large libration amplitudes and the speed at which the amplitudes grow. If the total range in Laplace angles, $\Delta\phi_1$ or $\Delta\phi_2$, exceeds a cut-off value K_1 over the integration time (T_{\max} , in years), then the time at which this occurs is recorded (T_{runaway}). A value $-1 + (T_{\text{runaway}}/T_{\max})^{-2}$ is added to the χ^2 value. All χ^2 values were also penalized by an additional term equal to $(\Delta\phi - V_i)^2$ if $\Delta\phi_i > V_i$ and to 0 if $\Delta\phi_i < V_i$ for specified angles V_i , $i = 1, 2$, in degrees and with $\Delta\phi = \phi_{\max} - \phi_{\min}$. This way, if the Laplace angles were well behaved enough not to run away, but either or both still grew in amplitude above specified values for each angle (V_1 and V_2), then a χ^2 penalty was assigned and the parameters were less likely to be accepted. We do not impose a direct eccentricity constraint. We report C_3 with $(T_{\max}, K_1, V_1, V_2) = (100 \text{ yr}, 170^\circ, 30^\circ, 50^\circ)$, for which the numbers are roughly based on the results of migration and DEMCMC results that had long-term libration (see Extended Data Table 2).

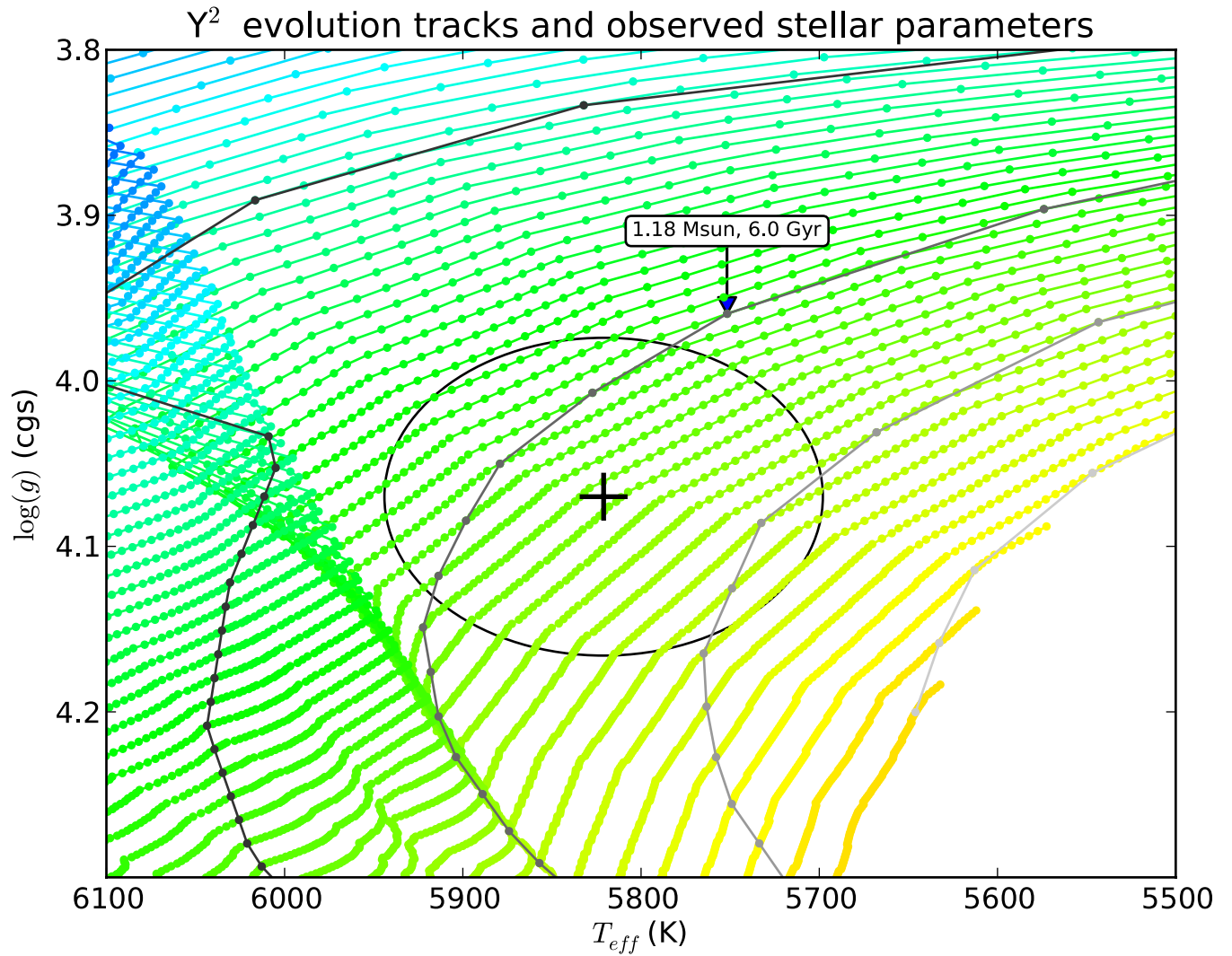
Running a similar stability check for C_3 as for C_1 by choosing 300 chains from the posterior distribution resulted in 100% of the parameter sets leading to stable behaviour lasting 10^7 years. Ten parameter sets were numerically integrated for 10^8 years, and 100% of those also lead to a system that survives with no close encounters. These results indicate that this method is effective at finding stable solutions. Comparing this to the stability results for C_1 , in which only 19% of solutions were stable for 10^8 years (as described above), our argument that resonance does encourage stability is strengthened. Nevertheless, this method cannot be guaranteed to reject all unstable systems (because they might pass this test) or to include all stable ones (because some systems could remain stable for a very long time, but have large changes in Laplace angle); see Extended Data Fig. 4. This posterior has lower eccentricities, but because we assume short-term resonance for this fit, we do not take it as our nominal fit.

Future observations. We predict future transit times and uncertainties by averaging the predicted transits from 152 solutions from the C_1 posterior that are stable for 10^7 years. We report transit times quarterly for 10 years, including over the Kepler observing window, in Supplementary Information.

Code availability. The code used for migration simulations is available as Supplementary Information. The code used to generate the TTV and photodynamic analyses is available upon request and will be made publicly available once further analyses have been completed.⁵⁰

- Vogt, S. S. *et al.* HRES: the high-resolution echelle spectrometer on the Keck 10-m Telescope. *Proc. SPIE* **2198**, 362–375 (1994).
- Coelho, P., Barbuy, B., Melendez, J., Schiavon, R. P. & Castilho, B. V. A library of high resolution synthetic stellar spectra from 300 nm to 1.8 μm with solar and α -enhanced composition. *Astron. Astrophys.* **443**, 735–746 (2005).
- Demarque, P., Woo, J.-H., Kim, Y.-C. & Yi, S. K. Y^2 isochrones with an improved core overshoot treatment. *Astrophys. J. Suppl. Ser.* **155**, 667–674 (2004).
- Morton, T. D. isochrones: stellar model grid package. *Astrophysics Source Code Library* ascl:1503.010, <http://ascl.net/1503.010> (2015).
- Girardi, L., Groenewegen, M. A. T., Hatziminaoglou, E. & da Costa, L. Star counts in the Galaxy. Simulating from very deep to very shallow photometric surveys with the TRILEGAL code. *Astron. Astrophys.* **436**, 895–915 (2005).
- Quillen, A. C. Three-body resonance overlap in closely spaced multiple-planet systems. *Mon. Not. R. Astron. Soc.* **418**, 1043–1054 (2011).
- Carter, J. A. *et al.* KOI-126: a triply eclipsing hierarchical triple with two low-mass stars. *Science* **331**, 562–565 (2011).
- ter Braak, C. J. F. *Genetic Algorithms and Markov Chain Monte Carlo: Differential Evolution Markov Chain Makes Bayesian Computing Easy*. Report No. 010404 (revised) <http://edepot.wur.nl/39477> (Biometris, 2005).
- Veras, D. & Armitage, P. J. The dynamics of two massive planets on inclined orbits. *Icarus* **172**, 349–371 (2004).
- Huber, D. *et al.* Stellar spin-orbit misalignment in a multiplanet system. *Science* **342**, 331–334 (2013).
- Masuda, K., Hirano, T., Taruya, A., Nagasawa, M. & Suto, Y. Characterization of the KOI-94 system with transit timing variation analysis: implication for the planet-planet eclipse. *Astrophys. J.* **778**, 185 (2013).
- Sing, D. K. Stellar limb-darkening coefficients for CoRoT and Kepler. *Astron. Astrophys.* **510**, A21 (2010).

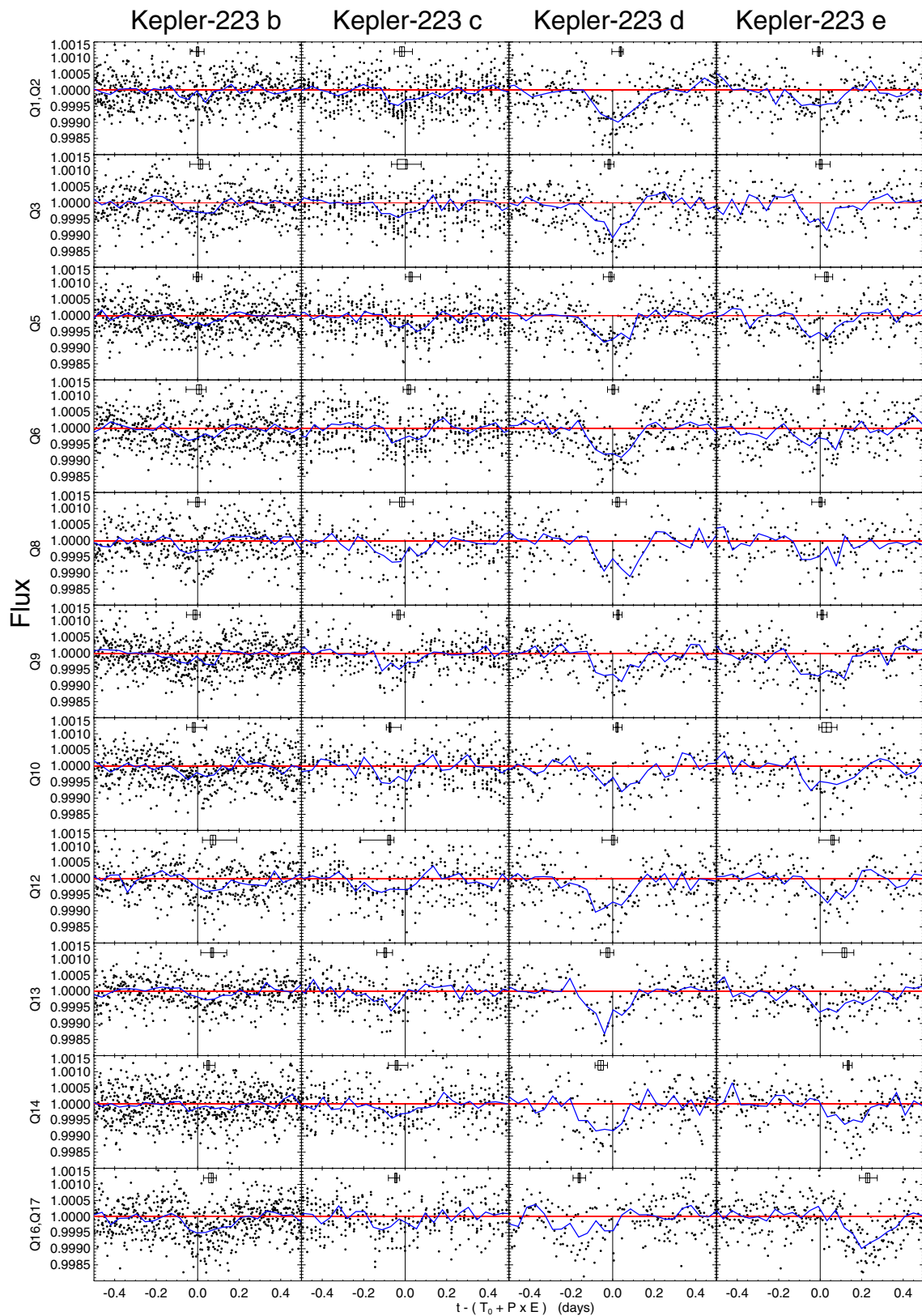
43. Southworth, J., Bruntt, H. & Buzasi, D. L. Eclipsing binaries observed with the WIRE satellite. II. β Aurigae and non-linear limb darkening in light curves. *Astron. Astrophys.* **467**, 1215–1226 (2007).
44. Southworth, J. Homogeneous studies of transiting extrasolar planets – I. Light-curve analyses. *Mon. Not. R. Astron. Soc.* **386**, 1644–1666 (2008).
45. Brown, T. M., Latham, D. W., Everett, M. E. & Esquerdo, G. A. *Kepler* input catalog: photometric calibration and stellar classification. *Astron. J.* **142**, 112 (2011).
46. Howell, S. B. *et al.* Kepler-21b: a $1.6 R_{\text{Earth}}$ planet transiting the bright oscillating F subgiant star HD 179070. *Astrophys. J.* **746**, 123 (2012).
47. Brown, T. M. *et al.* Las Cumbres Observatory Global Telescope network. *Publ. Astron. Soc. Pacif.* **125**, 1031–1055 (2013).
48. Howell, S. B., Everett, M. E., Sherry, W., Horch, E. & Ciardi, D. R. Speckle camera observations for the NASA *Kepler* mission follow-up program. *Astron. J.* **142**, 19 (2011).
49. Chambers, J. E. Mercury: a software package for orbital dynamics. *Astrophysics Source Code Library* ascl:1201.008, <http://ascl.net/1201.008> (2012).
50. Williams, J. P. & Cieza, L. A. Protoplanetary disks and their evolution. *Annu. Rev. Astron. Astrophys.* **49**, 67–117 (2011).
51. Lissauer, J. J. *et al.* Architecture and dynamics of *Kepler*'s candidate multiple transiting planet systems. *Astrophys. J. Suppl. Ser.* **197**, 8 (2011).
52. Batalha, N. M. *et al.* Planetary candidates observed by *Kepler*. III. Analysis of the first 16 months of data. *Astrophys. J. Suppl. Ser.* **204**, 24 (2013).



Extended Data Figure 1 | Spectroscopic fit of the Kepler-223 star.

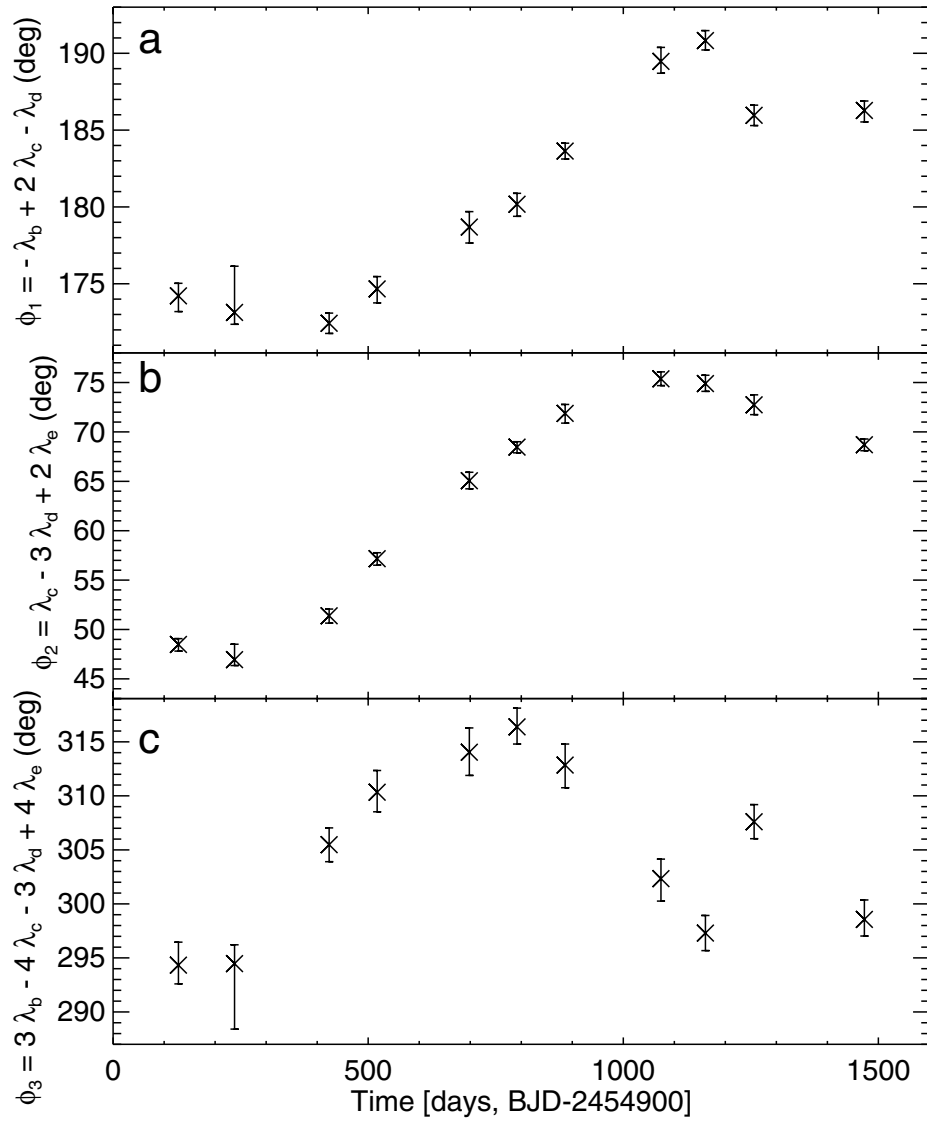
A fit to Yonsei–Yale (Y^2) evolution tracks (coloured lines) with 0.01-Gyr increments marked with filled circles. Colours correspond to mass with increments of $0.01M_{\odot}$ from $1.0M_{\odot}$ (orange) to $1.4M_{\odot}$ (darkest blue). Isochrones (grey lines) are over-plotted in 2-Gyr increments from 4 Gyr (darkest grey) to 10 Gyr (lightest grey) with filled circles every $0.01M_{\odot}$

increment. One point is labelled for reference ($M_{\text{sun}} = M_{\odot}$). The best-fit (T_{eff} , $\log(g)$) value (black cross) and an ellipse (black) whose semi-major axes indicate 1σ uncertainties of each parameter found from spectral matching are indicated. The stars in this area of parameter space have evolved off the main sequence.



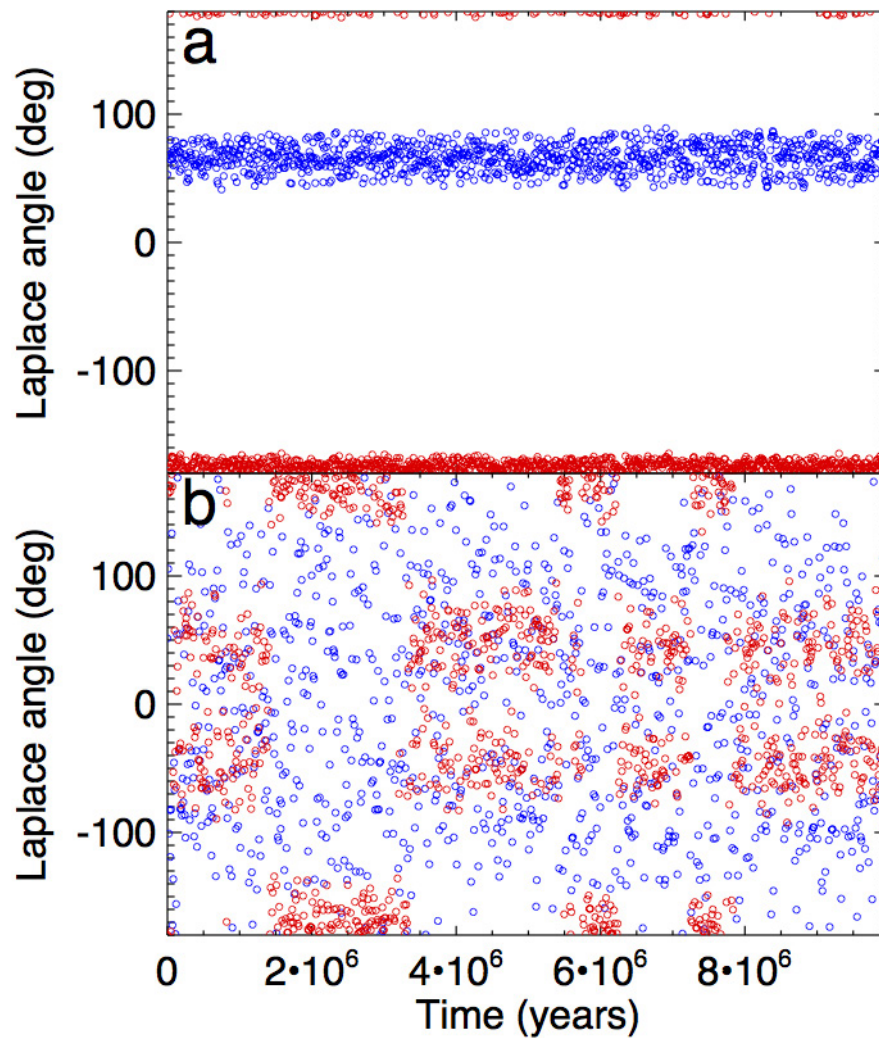
Extended Data Figure 2 | Long-cadence light curve for each planet, broken down by quarter (Q). Data (black filled circles) are binned via a moving average to give the blue curve, to reduce the scatter relative to the horizontal red line indicating no signal. Each panel is centred on the transit times predicted using the linear ephemeris (T_0 and P) of ref. 52 (vertical black lines), with the horizontal axis the time in days from the

Eth predicted transit time. The box-and-whisker error bars indicate the best-fit mid-transit time and 1σ and 3σ uncertainties based on $\Delta\chi^2=1$ and $\Delta\chi^2=9$. χ^2 values are computed by sliding an overall fit to the transit horizontally across the data and interpolating. Their offset relative to the linear ephemeris lines indicates the magnitudes of the TTVs.



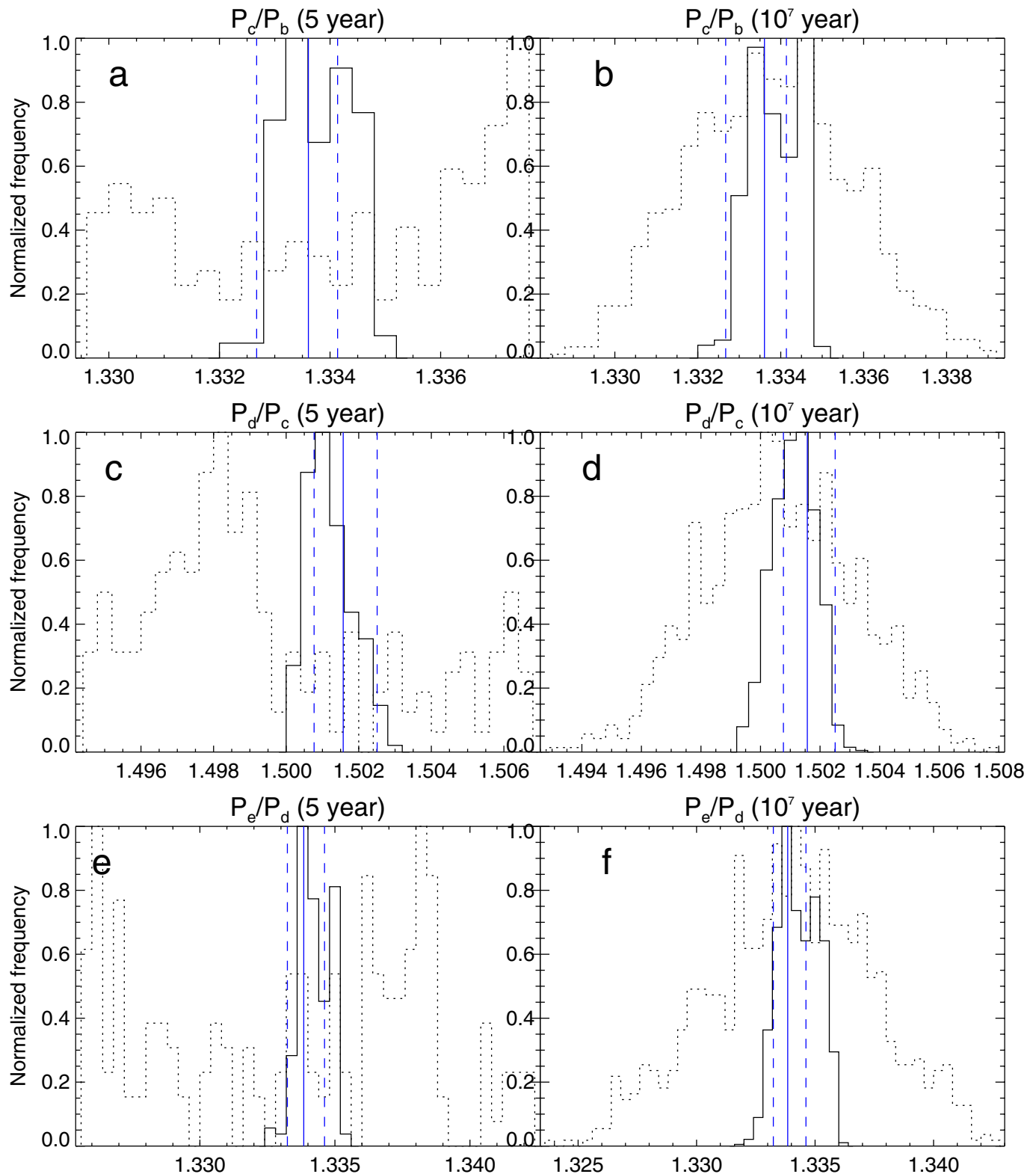
Extended Data Figure 3 | Laplace-angle librations detected by binning transits into quarters and assuming zero eccentricity. a–c, Error bars show 1σ uncertainties based on $\Delta\chi^2=1$. Almost a full libration cycle of all angles is observed in the $\sim 1,500$ -day observing window. The amplitude of oscillation in the four-body Laplace angle (ϕ_3 ; c) is similar in amplitude to

each of the individual Laplace angles (ϕ_1 , a; ϕ_2 , b). Because $\phi_3 = -3\phi_1 + 2\phi_2$, this amplitude could naively be expected to be much larger; however, ϕ_1 and ϕ_2 are closely related, owing to the four-body resonance of the Kepler-223 system, in contrast to two independent three-body resonances.



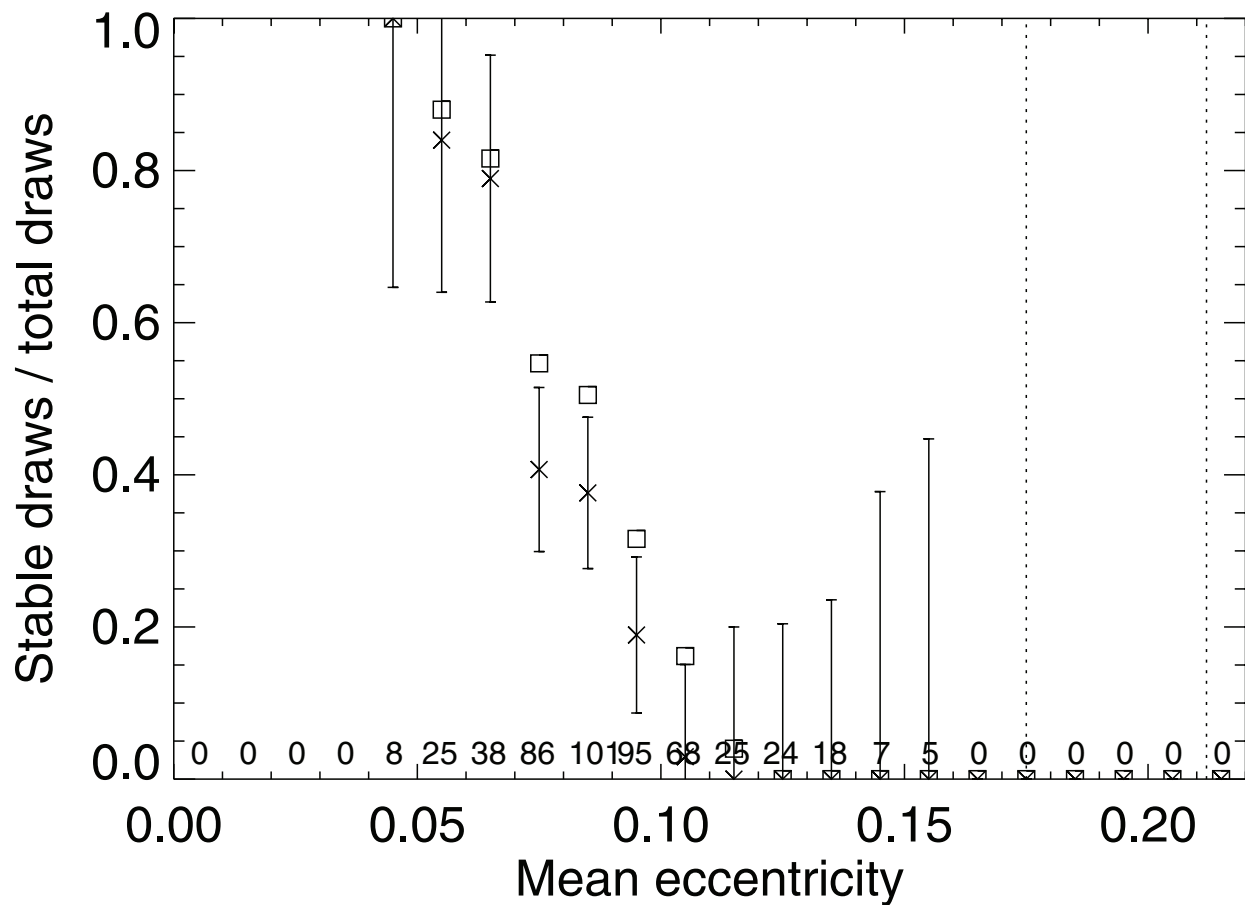
Extended Data Figure 4 | Variation in Laplace angles for two 10^7 -year-stable solutions. **a**, The librating Laplace angles (ϕ_1 , red; ϕ_2 , blue) for a solution from the \mathcal{C}_3 DEMCMC posterior. Laplace angles librate over the entire 10^7 years. The orbital-period distribution in Extended Data Fig. 5 uses this model. **b**, Another solution from \mathcal{C}_3 , in which the inner

Laplace angle (ϕ_1 ; red) librates near the observed value initially, but begins switching chaotically between three different libration centres. This is not uncommon in the \mathcal{C}_3 DEMCMC posterior. Despite the initial constraint on the outer Laplace angle (ϕ_2 ; blue), there are long periods of circulation with intermittent libration.



Extended Data Figure 5 | Orbital-period ratios of librating and non-librating solutions fitted to data. **a, c, e,** The distribution of osculating period ratios for each neighbouring planet pair (P_c/P_b , **a**; P_d/P_c , **c**; P_e/P_d , **e**) over a randomly selected 4-year window in the first 10^4 years for two 10^7 -year-stable parameter sets from the C_3 DEMCMC posterior solution. The dotted histogram represents a solution that showed

substantial periods of Laplace-angle circulation. The solid histogram represents a solution in which both ϕ_1 and ϕ_2 librate for 10^7 years. The blue vertical line indicates the empirical mean period; blue dashed vertical lines represent the highest and lowest quarter-to-quarter period measured. **b, d, f,** The same as **a, c, e**, but over the entire 10^7 -year interval.



Extended Data Figure 6 | System stability as a function of mean planetary eccentricity. The fraction of 500 random draws from the \mathcal{C}_1 posterior that survive for 10^7 years (crosses) and 10^6 years (squares) as a function of four-planet-mean eccentricity in bins of width 0.01. 1σ statistical uncertainties are included as vertical error bars on the

crosses. Dotted lines indicate the two eccentricity limits for the planets used in \mathcal{C}_1 : 0.175 (planets c and e) and 0.212 (planets b and d). Numbers represent the total number of draws in each eccentricity bin. The fraction of 10^7 -year-stable systems falls sharply and is consistent with zero well below the eccentricity cuts imposed by \mathcal{C}_1 .

Extended Data Table 1 | Mean Kepler-223 quarterly TTVs

$\hat{t} - 2454900$ (BJD)	-3σ	$-\sigma$	Best	$+\sigma$	$+3\sigma$
<i>Kepler-223b</i> : $P = 7.3840154$ days, $T_0 - 2454900$ (BJD) = 70.49489					
123.32662	-0.0354	-0.0058	-0.0006	0.0059	0.0316
239.02516	-0.0517	-0.0103	0.0137	0.0101	0.0423
416.51724	-0.0200	-0.0061	-0.0010	0.0062	0.0210
521.69775	-0.0628	-0.0123	0.0068	0.0113	0.0342
699.18988	-0.0470	-0.0088	-0.0010	0.0084	0.0370
797.79657	-0.0417	-0.0097	-0.0123	0.0088	0.0243
886.54260	-0.0343	-0.0072	-0.0187	0.0074	0.0617
1073.89526	-0.0500	-0.0118	0.0730	0.0140	0.1150
1162.64136	-0.0542	-0.0071	0.0692	0.0071	0.0708
1251.38745	-0.0217	-0.0062	0.0507	0.0065	0.0333
1458.46155	-0.0379	-0.0129	0.0659	0.0090	0.0241
<i>Kepler-223c</i> : $P = 9.8487130$ days, $T_0 - 2454900$ (BJD) = 71.37624					
116.10564	-0.0362	-0.0103	-0.0168	0.0133	0.0518
242.86336	-0.0683	-0.0405	0.0023	0.0077	0.0747
420.32413	-0.0254	-0.0077	0.0264	0.0076	0.0476
509.05453	-0.0266	-0.0077	0.0166	0.0090	0.0304
701.30371	-0.0585	-0.0121	-0.0155	0.0126	0.0535
790.03418	-0.0302	-0.0075	-0.0318	0.0084	0.0268
886.15869	-0.0173	-0.0046	-0.0737	0.0048	0.0537
1071.01367	-0.1404	-0.0078	-0.0766	0.0066	0.0226
1148.65283	-0.0411	-0.0067	-0.0959	0.0064	0.0349
1252.17163	-0.0392	-0.0067	-0.0418	0.0068	0.0548
1470.30054	-0.0361	-0.0056	-0.0449	0.0051	0.0179
<i>Kepler-223d</i> : $P = 14.7883997$ days, $T_0 - 2454900$ (BJD) = 109.76775					
132.10997	-0.0416	-0.0058	0.0376	0.0054	0.0134
248.65308	-0.0221	-0.0062	-0.0169	0.0063	0.0229
427.57138	-0.0351	-0.0084	-0.0099	0.0070	0.0169
519.49268	-0.0285	-0.0070	0.0035	0.0066	0.0245
711.54260	-0.0260	-0.0086	0.0240	0.0094	0.0420
800.18097	-0.0226	-0.0060	0.0256	0.0057	0.0194
898.66815	-0.0192	-0.0057	0.0212	0.0055	0.0238
1077.35193	-0.0530	-0.0080	0.0020	0.0077	0.0210
1169.50781	-0.0354	-0.0085	-0.0236	0.0093	0.0286
1271.27771	-0.0272	-0.0131	-0.0578	0.0132	0.0328
1483.43542	-0.0298	-0.0061	-0.1612	0.0057	0.0302
<i>Kepler-223e</i> : $P = 19.7213435$ days, $T_0 - 2454900$ (BJD) = 68.10686					
135.47421	-0.0303	-0.0060	-0.0067	0.0053	0.0187
238.21753	-0.0232	-0.0067	0.0022	0.0072	0.0458
433.78842	-0.0542	-0.0095	0.0302	0.0084	0.0298
524.27625	-0.0244	-0.0061	-0.0106	0.0063	0.0296
709.21222	-0.0432	-0.0071	0.0022	0.0065	0.0208
797.82037	-0.0240	-0.0060	0.0090	0.0061	0.0240
893.81256	-0.0357	-0.0216	0.0297	0.0242	0.0513
1079.88989	-0.0662	-0.0083	0.0602	0.0078	0.0308
1170.71301	-0.1067	-0.0118	0.1167	0.0110	0.0453
1263.01343	-0.0252	-0.0049	0.1352	0.0049	0.0188
1469.48169	-0.0393	-0.0097	0.2283	0.0100	0.0467

Transit times and TTVs (in days) for each planet found by binning the data quarterly and iteratively solving for transit shape as described in Methods. Mean transit time in the quarter is given in the first column followed by the measured TTV and uncertainties as described in Extended Data Fig. 2.

Extended Data Table 2 | Complete Kepler-223 parameters

Parameter Name (Unit)	Eccentricity Prior (\mathcal{C}_1)	Eccentricity Prior and Stability (\mathcal{C}_2)	Laplace Angle Constraint (\mathcal{C}_3)
<i>Stellar Parameters:</i>			
$R_*(R_\odot)$	$1.714^{+0.079}_{-0.165}$	$1.72^{+0.07}_{-0.14}$	$1.622^{+0.078}_{-0.070}$
$M_*(M_\odot)$	1.125 (fixed)	1.125 (fixed)	1.125 (fixed)
c_1	$0.54^{+0.11}_{-0.10}$	$0.54^{+0.10}_{-0.09}$	$0.57^{+0.11}_{-0.10}$
c_2	0.2 (fixed)	0.2 (fixed)	0.2 (fixed)
dilution	0.11202 (fixed)	0.11202 (fixed)	0.11202 (fixed)
<i>Kepler-223 b Parameters:</i>			
P (d)	$7.38454^{+0.00024}_{-0.00028}$	$7.38449^{+0.00022}_{-0.00022}$	$7.38453^{+0.00024}_{-0.00024}$
T_0 (BJD-2454900)	$801.5145^{+0.0044}_{-0.0047}$	$801.5155^{+0.0044}_{-0.0046}$	$801.5133^{+0.0042}_{-0.0045}$
$e \cdot \cos(\omega)$	$0.057^{+0.034}_{-0.031}$	$0.054^{+0.022}_{-0.022}$	$0.035^{+0.014}_{-0.016}$
$e \cdot \sin(\omega)$	$0.052^{+0.026}_{-0.135}$	$0.047^{+0.020}_{-0.039}$	$-0.004^{+0.029}_{-0.034}$
$ i - 90 $ ($^\circ$)	$0.0^{+1.7}_{-0.0}$	$0.0^{+1.8}_{-0.0}$	$0.0^{+1.4}_{-0.0}$
Ω ($^\circ$)	0.0 (fixed)	0.0 (fixed)	0.0 (fixed)
M/M_*	$0.0000196^{+0.0000034}_{-0.0000031}$	$0.0000221^{+0.0000032}_{-0.0000031}$	$0.0000201^{+0.0000027}_{-0.0000026}$
R/R_*	$0.01596^{+0.00053}_{-0.00053}$	$0.01597^{+0.00055}_{-0.00054}$	$0.01584^{+0.00052}_{-0.00053}$
<i>Kepler-223 c Parameters:</i>			
P (d)	$9.84584^{+0.00085}_{-0.00053}$	$9.84564^{+0.00052}_{-0.00051}$	$9.84613^{+0.00046}_{-0.00045}$
T_0 (BJD-2454900)	$800.1461^{+0.0049}_{-0.0040}$	$800.1459^{+0.0050}_{-0.0039}$	$800.1489^{+0.0061}_{-0.0047}$
$e \cdot \cos(\omega)$	$0.030^{+0.050}_{-0.047}$	$0.029^{+0.041}_{-0.038}$	$-0.010^{+0.019}_{-0.022}$
$e \cdot \sin(\omega)$	$0.134^{+0.027}_{-0.156}$	$0.139^{+0.021}_{-0.050}$	$0.060^{+0.033}_{-0.038}$
$ i - 90 $ ($^\circ$)	$0.0^{+1.4}_{-0.0}$	$0.0^{+1.3}_{-0.0}$	$0.0^{+1.5}_{-0.0}$
Ω ($^\circ$)	0.0 (fixed)	0.0 (fixed)	0.0 (fixed)
M/M_*	$0.0000157^{+0.0000048}_{-0.0000038}$	$0.0000152^{+0.0000048}_{-0.0000033}$	$0.0000189^{+0.0000032}_{-0.0000033}$
R/R_*	$0.01847^{+0.00055}_{-0.00056}$	$0.01842^{+0.00053}_{-0.00053}$	$0.01833^{+0.00056}_{-0.00057}$
<i>Kepler-223 d Parameters:</i>			
P (d)	$14.78881^{+0.00049}_{-0.00040}$	$14.78869^{+0.00030}_{-0.00027}$	$14.78862^{+0.00025}_{-0.00024}$
T_0 (BJD-2454900)	$804.8502^{+0.0022}_{-0.0023}$	$804.8504^{+0.0023}_{-0.0024}$	$804.8492^{+0.0022}_{-0.0023}$
$e \cdot \cos(\omega)$	$0.020^{+0.031}_{-0.030}$	$0.020^{+0.026}_{-0.024}$	$0.000^{+0.011}_{-0.013}$
$e \cdot \sin(\omega)$	$0.017^{+0.023}_{-0.076}$	$0.010^{+0.020}_{-0.032}$	$-0.001^{+0.015}_{-0.021}$
$ i - 90 $ ($^\circ$)	$2.02^{+0.29}_{-0.52}$	$2.06^{+0.26}_{-0.32}$	$1.68^{+0.30}_{-0.29}$
Ω ($^\circ$)	0.0 (fixed)	0.0 (fixed)	0.0 (fixed)
M/M_*	$0.0000203^{+0.0000040}_{-0.0000039}$	$0.0000240^{+0.0000039}_{-0.0000035}$	$0.0000225^{+0.0000032}_{-0.0000032}$
R/R_*	$0.02791^{+0.00056}_{-0.00064}$	$0.02800^{+0.00052}_{-0.00059}$	$0.02756^{+0.00053}_{-0.00058}$
<i>Kepler-223 e Parameters:</i>			
P (d)	$19.72553^{+0.00067}_{-0.00071}$	$19.72567^{+0.00055}_{-0.00054}$	$19.72568^{+0.00054}_{-0.00048}$
T_0 (BJD-2454900)	$817.5231^{+0.0055}_{-0.0048}$	$817.5237^{+0.0055}_{-0.0051}$	$817.5231^{+0.0053}_{-0.0046}$
$e \cdot \cos(\omega)$	$0.017^{+0.042}_{-0.033}$	$0.017^{+0.026}_{-0.024}$	$0.013^{+0.014}_{-0.014}$
$e \cdot \sin(\omega)$	$0.045^{+0.032}_{-0.077}$	$0.039^{+0.023}_{-0.032}$	$0.033^{+0.016}_{-0.023}$
$ i - 90 $ ($^\circ$)	$1.95^{+0.25}_{-0.45}$	$2.00^{+0.21}_{-0.27}$	$1.69^{+0.25}_{-0.24}$
Ω ($^\circ$)	0.0 (fixed)	0.0 (fixed)	0.0 (fixed)
M/M_*	$0.0000102^{+0.0000044}_{-0.0000042}$	$0.0000145^{+0.0000039}_{-0.0000036}$	$0.0000130^{+0.0000031}_{-0.0000029}$
R/R_*	$0.02450^{+0.00076}_{-0.00077}$	$0.02466^{+0.00074}_{-0.00076}$	$0.02421^{+0.00069}_{-0.00068}$

DEMC MC posterior probability median values and 68% confidence intervals for all model parameters at $T_{\text{epoch}} = 800.0$ (BJD - 2,454,900). Three parameter sets are given with fixed stellar mass: (1) DEMCMC results with eccentricity constraint \mathcal{C}_1 ; (2) the subset of the \mathcal{C}_1 DEMCMC results that retain only those solutions that are stable for 10^6 years (\mathcal{C}_2); and (3) Laplace-angle constraint \mathcal{C}_3 and fixed $\Omega_i = 0$ for $i = b, c, d, e$.

Extended Data Table 3 | Best-fit Kepler-223 initial conditions

Planet	Period (d)	T_0 (BJD-2454900)	e	i ($^\circ$)	Ω ($^\circ$)	ω ($^\circ$)	Mass (M_{Jup})	Radius (R_p/R_*)
b	7.384720365879194	801.516262774051825	0.105758145660053	90.701847866139545	0.0	62.597372675420416	0.022730704097050	0.015954404145479
c	9.845453934132928	800.146170501596430	0.172729064427036	90.301811036839879	0.0	85.015828120049491	0.017312231285438	0.018346434846992
d	14.788902636701252	804.851045349929109	0.037330052890247	92.189693102657941	0.0	76.465729705828863	0.019623186719198	0.027674878130791
e	19.726218957815664	817.521944355066694	0.051464531998599	92.056638725826986	0.0	111.706814565803512	0.009576406850388	0.024759859857039
Stellar Parameters:								
	1.125 M_* (M_\odot)		R_* (R_\odot):	1.744528317200141	c_1 :	0.479330549583184	c_2 : 0.2	dilute: 0.11202
b	7.384583733215798	801.513943095097261	0.061453702027857	91.105539095271382	0.0	37.604238003695137	0.020503806935496	0.015793288256059
c	9.845639757204141	800.144691508369419	0.112391047984129	91.085286013475226	0.0	86.059011138583742	0.019192688432573	0.0186099595659302
d	14.788880252356291	804.849755312464254	0.026604678672708	91.966288309512123	0.0	58.807213313926120	0.025560722351934	0.028232411829371
e	19.725687523818440	817.519383441790524	0.060783217179960	91.806556478578258	0.0	76.156009027159996	0.015467248730564	0.024265426463497
Stellar Parameters:								
	1.125 M_* (M_\odot)		R_* (R_\odot):	1.683974231305496	c_1 :	0.532243950638929	c_2 : 0.2	dilute: 0.11202

Best-fit initial planet conditions found by DEMCMC under \mathcal{C}_1 (top) and \mathcal{C}_3 (bottom) constraints at $T_{\text{epoch}}=800.0$ (BJD $-2,454,900$) with $\chi^2=746,480$ and $\chi^2=746,489$, respectively. M_{Jup} , mass of Jupiter.

A high-temperature ferromagnetic topological insulating phase by proximity coupling

Ferhat Katmis^{1,2,3*}, Valeria Lauter^{4*}, Flavio S. Nogueira^{5,6}, Badih A. Assaf^{7,8}, Michelle E. Jamer⁷, Peng Wei^{1,2,3}, Biswarup Satpati⁹, John W. Freeland¹⁰, Ilya Eremin⁵, Don Heiman⁷, Pablo Jarillo-Herrero¹ & Jagadeesh S. Moodera^{1,2,3}

Topological insulators are insulating materials that display conducting surface states protected by time-reversal symmetry^{1,2}, wherein electron spins are locked to their momentum. This unique property opens up new opportunities for creating next-generation electronic, spintronic and quantum computation devices^{3–5}. Introducing ferromagnetic order into a topological insulator system without compromising its distinctive quantum coherent features could lead to the realization of several predicted physical phenomena^{6,7}. In particular, achieving robust long-range magnetic order at the surface of the topological insulator at specific locations without introducing spin-scattering centres could open up new possibilities for devices. Here we use spin-polarized neutron reflectivity experiments to demonstrate topologically enhanced interface magnetism by coupling a ferromagnetic insulator (EuS) to a topological insulator (Bi₂Se₃) in a bilayer system. This interfacial ferromagnetism persists up to room temperature, even though the ferromagnetic insulator is known to order ferromagnetically only at low temperatures (<17 K). The magnetism induced at the interface resulting from the large spin-orbit interaction and the spin-momentum locking of the topological insulator surface greatly enhances the magnetic ordering (Curie) temperature of this bilayer system. The ferromagnetism extends ~2 nm into the Bi₂Se₃ from the interface. Owing to the short-range nature of the ferromagnetic exchange interaction, the time-reversal symmetry is broken only near the surface of a topological insulator, while leaving its bulk states unaffected. The topological magneto-electric response originating in such an engineered topological insulator^{2,8} could allow efficient manipulation of the magnetization dynamics by an electric field, providing an energy-efficient topological control mechanism for future spin-based technologies.

Realizing a ferromagnetic surface state in a topological insulator (TI) is predicted to allow several prominent phenomena to emerge, such as the interfacial magneto-electric effect⁹, the electric-field-induced image magnetic monopole^{1,2}, and Majorana fermions¹⁰. To achieve this goal, we need to introduce ferromagnetism at the surface of the TI while leaving its bulk properties unchanged¹¹. The magnetic proximity by interfacial exchange coupling¹² allows us to avoid the introduction of defects and reliably to separate bulk from surface state effects, which are advantages over magnetic bulk doping¹³ or surface doping by magnetic adatoms¹⁴. In practice, the current technology of inducing magnetism in TI is confined to low temperatures. Furthermore, there is a lack of experimental evidence and of detailed understanding of the proximity-induced magnetism, restricting its potential for applications. A key requirement for useful applications is the generation of room-temperature ferromagnetism in the TI¹⁵.

We engineered hybrid heterostructures of a TI (Bi₂Se₃) combined with a ferromagnetic insulator (FMI) (EuS) in bilayers grown with well defined atomically sharp interfaces and crystalline orientation. We elucidate the interactions at the interface between TI and FMI using depth-sensitive polarized neutron reflectometry (PNR). Using PNR we directly observe the emergence of a ferromagnetic state in the top two quintuple layers (QL, where 1 QL ≈ 0.96 nm) of Bi₂Se₃ near the TI-FMI interface. This ferromagnetic state in the TI persists up to temperatures larger than 300 K, far above the Curie temperature (above which the material normally loses its magnetism) of ~17 K of bulk EuS. The observation of an anomalous Hall effect provided additional evidence of a perpendicular moment in the TI that persists to high temperatures. Our findings demonstrate that such hybrid heterostructures can be implemented to achieve a robust and uniform surface magnetism in the TI, with high-quality coverage over a large area, which is a fundamental step towards device design.

We grew hybrid Bi₂Se₃-EuS bilayer structures by molecular beam epitaxy on sapphire (Al₂O₃(0001)) substrates. The X-ray diffraction (XRD) scans along the growth direction for different bilayer configurations and for a Bi₂Se₃ film alone are shown in Fig. 1a. The plots clearly show EuS(111) peaks for EuS layers 2–10 nm thick on different Bi₂Se₃ layer thicknesses. The microstructure of the layers is visible in the cross-sectional high-resolution transmission electron microscopy (TEM) images in Fig. 1b and c, showing that EuS and Bi₂Se₃ are both coherently aligned along the (111) and the (0001) directions, respectively, and an atomically sharp interface is formed at the boundary. The QL structure of Bi₂Se₃ is clearly resolved, whereas individual atomic planes are seen for EuS (in Fig. 1c). XRD and high-resolution TEM thus confirm the formation of highly ordered heterostructures for all samples. The interface quality was further confirmed with soft X-ray absorption spectroscopy, simultaneously using surface- and bulk-sensitive techniques, which indicate a sharp electronic interface between EuS and Bi₂Se₃ (see Extended Data Fig. 1 for details).

Thin films of EuS on Si substrates generally favour in-plane magnetic anisotropy even down to a 1-nm thickness¹⁶. However, we find that in close proximity with Bi₂Se₃, the strong spin-orbit coupling modifies the in-plane anisotropy^{17,18}, thereby leading to an out-of-plane magnetic moment in the TI surface state¹². We found that all films display a remanent moment in both in-plane and out-of-plane geometries (see Extended Data Fig. 2), where the out-of-plane remanence is a clear signature of the normally in-plane moment being tilted out-of-plane. This tilting was found to increase for thinner TI films, suggesting an enhanced out-of-plane preference. To understand the nature of the anisotropy, superconducting quantum interference device (SQUID) magnetometry of different thicknesses of EuS and

¹Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ²Francis Bitter Magnet Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ³Plasma Science and Fusion Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ⁴Quantum Condensed Matter Division, Neutron Sciences Directorate, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA. ⁵Institut fuer Theoretische Physik III, Ruhr-Universitaet Bochum, D-44801 Bochum, Germany. ⁶Institute for Theoretical Solid State Physics, Institut fuer Festkoerper- und Werkstofforschung, Dresden, D-01069 Dresden, Germany. ⁷Department of Physics, Northeastern University, Boston, Massachusetts 02115, USA. ⁸Département de Physique, Ecole Normale Supérieure, Centre National de la Recherche Scientifique, Paris Sciences et Lettres Research University, Paris 75005, France. ⁹Saha Institute of Nuclear Physics, 1/AF Bidhannagar, Kolkata 64, India. ¹⁰Advanced Photon Source, Argonne National Laboratory, Argonne, Illinois 60439, USA.

*These authors contributed equally to this work.

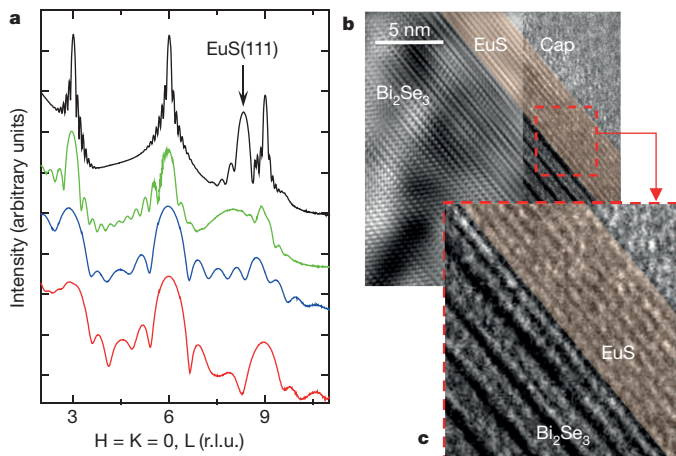


Figure 1 | XRD and high-resolution TEM of Bi₂Se₃-EuS bilayers.

a, XRD line scans along the L-direction (growth direction) for the bilayers with different Bi₂Se₃ and EuS thicknesses show that [111]-EuS is parallel to [0001]-Bi₂Se₃. Bi₂Se₃-EuS bilayers of thickness 30 QL/10 nm (black), 10 QL/2 nm (green) and 5 QL/1 nm (blue) are shown. In addition Bi₂Se₃ (red) without EuS of thickness 5 QL is shown for comparison. Well defined Kiessig fringes are an indication of good correlation between the top surface of EuS and the interface of EuS and Bi₂Se₃. The graph is base-10 logarithmic on the y axis (scattered intensity) and linear on the x axis, where H, K, and L stand for the Miller indices in units of the reciprocal lattice (r.l.u.) of the Bi₂Se₃(0001) surface. **b**, Fourier-filtered cross-sectional high-resolution TEM image for the Bi₂Se₃-EuS interface. Bilayers are protected with an amorphous Al₂O₃ cap layer (Cap). **c**, Expanded image of the Bi₂Se₃-EuS interface, showing the defect-free and cluster-free atomically sharp bilayer interface.

Bi₂Se₃ bilayer combinations were compared. To quantify this change in magnetic anisotropy as a function of bilayer parameters we measured the in-plane anisotropy constant $K_1 = \frac{1}{2}\mu_0 H_A M_{\text{sat}}$ (ref. 19) and the remanence ratio M_0/M_{sat} . Here, M_0 is the magnetization at zero applied field, M_{sat} is the saturation magnetization, and H_A is the saturation field in the out-of-plane direction. Figure 2 plots K_1 and M_0/M_{sat} for various Bi₂Se₃ and EuS thicknesses. Interestingly, the in-plane anisotropy constant decreases systematically as the thickness of Bi₂Se₃ is reduced (Fig. 2a). This is accompanied by a decrease in the remanence ratio in the in-plane direction as expected (Fig. 2b). The same decrease in the remanence ratio is observed when the EuS thickness is decreased (Fig. 2c). In contrast, we found that the out-of-plane remanence ratio is remarkably unaffected when either thickness is changed, suggesting that the out-of-plane component is an interface effect. We measured an average remanence of $\sim 6\%$ in the out-of-plane direction.

From the magnetization studies above and also from a previous study¹², we expect the Bi₂Se₃ layer to become magnetic and spin-polarized owing to the exchange coupling with the adjacent FMI layer. The experimentally observed out-of-plane component is necessary for splitting the bands in the TI and breaking the time-reversal symmetry at the interfacial region adjacent to the FMI^{1,2,8,20}. This can be regarded as a consequence of the strong spin-orbit coupling, which leads to the locking of spin to momentum, with the in-plane fluctuations contributing to the Berry phase²¹⁻²³.

To understand the exchange interaction better and to explore the depth profile of the magnetism at the interface directly, we used a depth-sensitive PNR technique²⁴. The depth profiles of the nuclear and magnetic scattering length densities (NSLD and MSLD; see Extended Data Fig. 3 for details) correspond to the depth profile of the chemical and in-plane magnetization vector distributions, respectively. PNR measurements were carried out on the bilayers with EuS thickness fixed at 5 nm and Bi₂Se₃ thickness varied (5 QL, 10 QL and 20 QL). Figure 3b shows results of R^+ and R^- reflectivity, where the superscript plus (or minus) signs indicate neutrons with spin parallel (or antiparallel) to the direction of the applied magnetic field, for samples measured at 5 K with the in-plane magnetic field $H_{\text{ext}} = 1$ T after the samples were

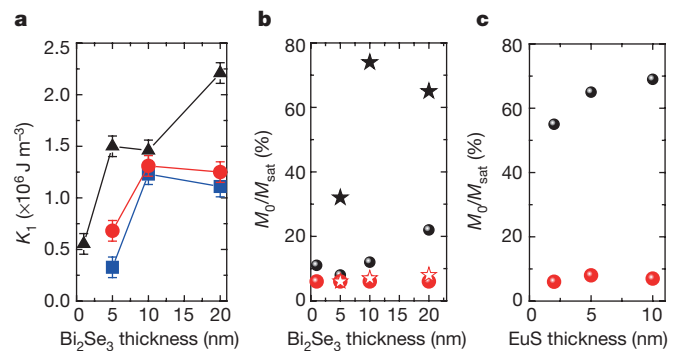


Figure 2 | SQUID magnetometry measurements for different Bi₂Se₃-EuS bilayers.

a, Magnetic anisotropy, K_1 , for various bilayer thickness combinations (the EuS thickness of 10 nm is shown in blue, 5 nm in red and 1 nm in black). The error bars come from the computation of M_{sat} from SQUID measurements. **b** and **c**, In-plane (IP, black) and out-of-plane (OP, red) remanence ratios M_0/M_{sat} for different bilayer samples are extracted from $M(H)$ loops for each sample. **b**, The EuS thickness is fixed at 1 nm (samples shown as circle symbols) and 5 nm (samples shown as star symbols) and Bi₂Se₃ ranges from 1 QL up to 20 QL. **c**, The Bi₂Se₃ thickness (20 QL) is fixed and EuS thicknesses range from 2 to 10 nm.

cooled at zero magnetic field. The NSLD and MSLD depth profiles were obtained from a simultaneous fit to the data and plotted as functions of the depth from the surface and shown in Fig. 3c for the sample with 20 QL Bi₂Se₃ (see Extended Data Fig. 3 for samples with 5 QL and 10 QL). PNR reveals a sharp interface between the EuS and Bi₂Se₃ layers over the whole lateral size of the sample with average roughness ~ 0.2 nm, as also observed by cross-sectional TEM. The absorption scattering length density (ASLD) depth profile (Fig. 3c), which is the signature of solely Eu atoms, stops at a certain depth, demonstrating that no Eu atoms are detected in the Bi₂Se₃ layer.

Remarkably, the magnetization profile MSLD shows a magnetization of 240 electromagnetic units (emu) cm^{-3} and 34 emu cm^{-3} in the first and second QL of Bi₂Se₃, respectively (marked with red arrows in Fig. 3c, where the Eu absorption length is shown as a blue curve), penetrating into Bi₂Se₃ beyond the EuS-Bi₂Se₃ interface. There is a concurrent reduction in the EuS magnetization observed near the interface, where 1.5 nm of the EuS layer has the moment reduced to 2.5 Bohr magnetons (μ_B) per Eu²⁺ (blue arrow in Fig. 3c), which is only about $\sim 36\%$ of the maximum $7\mu_B$ per Eu²⁺ in the bulk of the EuS. Given that the NSLD depth profile of the EuS layer is uniform and no changes are detected in the structural and chemical composition in this interfacial EuS layer, we attribute the reduced in-plane magnetization in this thin interfacial EuS layer to a canting of the Eu magnetization vector towards the out-of-plane direction. Since the out-of-plane component of the magnetization vector is parallel to the momentum transfer Q , it is thus not responsive in PNR²⁵. This is consistent with the observation of the out-of-plane magnetization component in SQUID measurements (Fig. 2). The sample's reflectivity below its critical edge, called the total reflection region, is unity if there is no absorption²⁶. The inset of Fig. 3b shows a magnified region of the total reflection. Here, the impact of the Eu absorption cross-section on the PNR reflectivity results in a striking feature in the total reflection region of R^+ and R^- , which is very sensitive to the depth profile of the Eu atoms. The ASLD profile in Fig. 3c shows a sharp interface between the EuS and Bi₂Se₃ layers as well, and confirms that Eu atoms are not present in the Bi₂Se₃ layer.

We investigated the magnetization behaviour of bilayers with PNR experiments above the Curie temperature (T_C) of EuS. We discovered that the heterostructures exhibit ferromagnetic behaviour even at 300 K. Figure 4a displays the spin-asymmetry (SA) ratio for the 10 QL Bi₂Se₃ sample measured at 50 K, 75 K, 120 K and 300 K. This striking observation of room-temperature ferromagnetism demonstrates that the non-zero magnetization present in the 2 QL Bi₂Se₃ interfacial layer also penetrates into the EuS layer (Fig. 4b), thus stabilizing

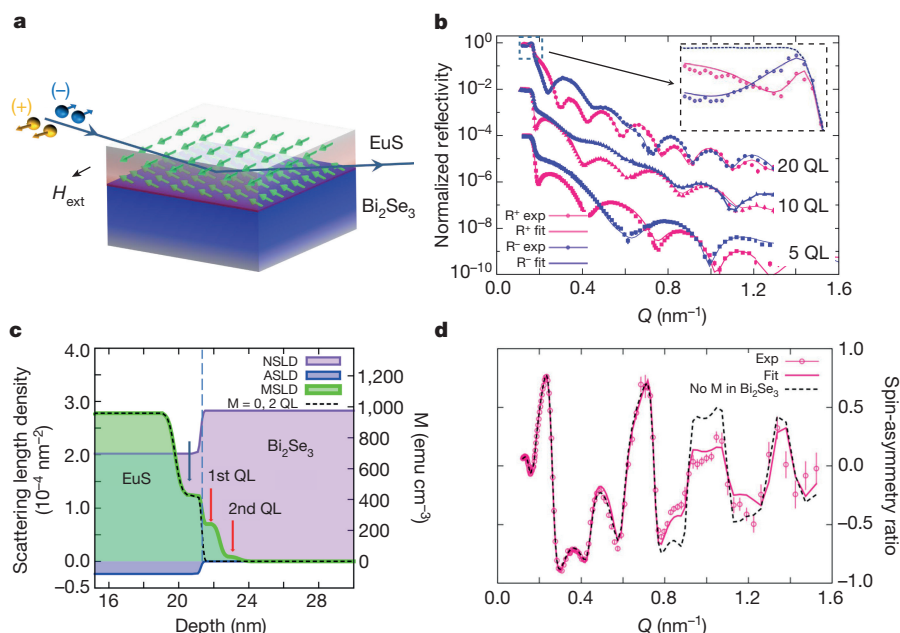


Figure 3 | Polarized neutron reflectivity results for different Bi_2Se_3 -EuS bilayers. **a**, Schematic of the PNR experimental set-up for Bi_2Se_3 -EuS bilayer films. **b**, Measured (symbols) and fitted (solid lines) reflectivity curves for spin-up (R^+) and spin-down (R^-) neutron spin-states (logarithmic-linear scale) shown as a function of momentum transfer $Q = 4\pi\sin(\theta)/\lambda$, where θ is the incident angle and λ is the neutron wavelength. The inset is an expanded view of the reflectivity below its critical edge, where the reflectivity is sensitive to the distribution of the Eu atoms owing to their absorption cross-section and their magnetic moment. The error bars represent one standard deviation. **c**, PNR nuclear (NSLD, in pink), magnetic (MSLD, in green) and absorption (ASLD, in blue) scattering length density profiles, measured for the 20 QL sample at 5 K

with an external in-plane magnetic field of 1 T and presented as a function of the distance from the sample surface. The magnetization measured inside the Bi_2Se_3 layer is marked with red arrows, and the reduction of the in-plane component of EuS at the interface caused by a canting of the Eu magnetization vector towards the OP direction is marked with a blue arrow. The scale on the right-hand-side shows magnetization M . **d**, PNR spin-asymmetry (SA) ratio $SA = (R^+ - R^-)/(R^+ + R^-)$ obtained from the experimental and fitted reflectivities in **b**. The fit with zero magnetization ($M = 0$ in 2 QL) in the Bi_2Se_3 layer (black dashed line in **d** obtained with the corresponding MSLD profile also shown with black dashed line in **c**) has a large deviation from the experimental data. The error bars represent one standard deviation.

the magnetization well above its T_C . The magnitude of the Bi_2Se_3 moment as a function of temperature is shown in Fig. 4c: although the TI moment is reduced by an order of magnitude at 120 K compared to its 5 K value, and by another factor of two or more at room temperature, it remains nevertheless substantial. No magnetization was detected above ~ 50 K in the pure EuS film when measured under the same experimental conditions as in bilayer films (see Extended Data Fig. 3 for details).

Thus we have successfully established ferromagnetic order at the surface of epitaxial Bi_2Se_3 films using internal exchange coupling through proximity with the ferromagnetic insulator EuS. PNR provides direct evidence that Bi_2Se_3 -EuS heterostructures exhibit proximity-induced interfacial magnetization in the top 2 QL (~ 2 nm) layer of Bi_2Se_3 . Thus

PNR enables us efficiently to distinguish the magnetic TI surface states from the trivial bulk states. We show that such effects originate through exchange interaction, without structural perturbation at the interface. Our PNR, magnetization, and transport studies reveal that magnetic moment persists in the TI at temperatures far above the Curie temperature of the FMI, signifying a robust topological magnetic state of the bilayer system. Owing to the short-range nature of this ferromagnetic exchange interaction, we are able to locally break the time-reversal symmetry on the surface of the TI while leaving its bulk states unaffected. Finally, the results reported here pave the way for a new class of spin-based electronics driven by gapped Dirac surface states. For instance, high-temperature ferromagnetism in gated FMI-TI-FMI structures may allow for the stabilization and

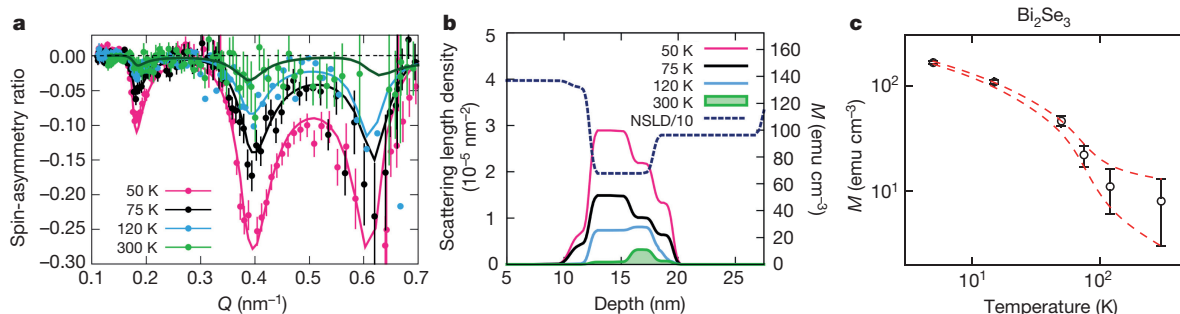


Figure 4 | Ferromagnetic order in Bi_2Se_3 -EuS bilayer samples. **a**, PNR measurements of a Bi_2Se_3 10 QL EuS bilayer film. The measured SA at different temperatures and model fits (shown with solid lines), where the sample was cooled at zero magnetic field then measured at 50 K, 75 K, 120 K and 300 K while warming. The error bars represent one standard deviation. **b**, Chemical (NSLD, dashed line) and magnetic (MSLD) depth

profiles at different temperatures (solid lines for temperatures 50 K, 75 K and 120 K and green shading for 300 K) are presented as a function of the distance from the sample surface. The scale on the right-hand side shows magnetization M . **c**, PNR-derived magnetization of Bi_2Se_3 as a function of temperature (logarithmic-logarithmic scale). The error bars indicate the confidence interval.

observation of the topological magnetoelectric effect. This would lead to unprecedented control of spin and charge carriers by means of a topological magnetoelectric bias mechanism.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 April 2015; accepted 3 March 2016.

Published online 9 May 2016.

- Hasan, M. Z. & Kane, C. L. Colloquium: Topological insulators. *Rev. Mod. Phys.* **82**, 3045–3067 (2010).
- Qi, X.-L. & Zhang, S.-C. Topological insulators and superconductors. *Rev. Mod. Phys.* **83**, 1057–1110 (2011).
- Fu, L. & Kane, C. L. Superconducting proximity effect and Majorana fermions at the surface of a topological insulator. *Phys. Rev. Lett.* **100**, 096407 (2008).
- Akhmerov, A., Nilsson, J. & Beenakker, C. Electrically detected interferometry of Majorana fermions in a topological insulator. *Phys. Rev. Lett.* **102**, 216404 (2009).
- Ferreira, G. J. & Loss, D. Magnetically defined qubits on 3D topological insulators. *Phys. Rev. Lett.* **111**, 106802 (2013).
- Chang, C. Z. *et al.* Experimental observation of the quantum anomalous Hall effect in a magnetic topological insulator. *Science* **340**, 167–170 (2013).
- Checkelsky, J. G. *et al.* Trajectory of the anomalous Hall effect towards the quantized state in a ferromagnetic topological insulator. *Nature Phys.* **10**, 731–736 (2014).
- Qi, X.-L., Hughes, T. L. & Zhang, S.-C. Topological field theory of time-reversal invariant insulators. *Phys. Rev. B* **78**, 195424 (2008); erratum *Phys. Rev. B* **81**, 159901 (2010).
- Essin, A., Moore, J. & Vanderbilt, D. Magnetoelectric polarizability and axion electrodynamics in crystalline insulators. *Phys. Rev. Lett.* **102**, 146805 (2009); erratum *Phys. Rev. Lett.* **103**, 259902 (2009).
- Nadj-Perge, S. *et al.* Observation of Majorana fermions in ferromagnetic atomic chains on a superconductor. *Science* **346**, 602–607 (2014).
- Scholz, M. R. *et al.* Tolerance of topological surface states towards magnetic moments: Fe on Bi₂Se₃. *Phys. Rev. Lett.* **108**, 256810 (2012).
- Wei, P. *et al.* Exchange-coupling-induced symmetry breaking in topological insulators. *Phys. Rev. Lett.* **110**, 186807 (2013).
- Chen, Y. L. *et al.* Massive Dirac fermion on the surface of a magnetically doped topological insulator. *Science* **329**, 659–662 (2010).
- Vobornik, I. *et al.* Magnetic proximity effect as a pathway to spintronic applications of topological insulators. *Nano Lett.* **11**, 4079–4082 (2011).
- Mellnik, A. R. *et al.* Spin-transfer torque generated by a topological insulator. *Nature* **511**, 449–451 (2014).
- Miao, G.-X. & Moodera, J. S. Controlling magnetic switching properties of EuS for constructing double spin filter magnetic tunnel junctions. *Appl. Phys. Lett.* **94**, 182504 (2009).
- Chappert, C. & Bruno, P. Magnetic anisotropy in metallic ultrathin films and related experiments on cobalt films. *J. Appl. Phys.* **64**, 5736 (1988).
- Semenov, Y. G., Duan, X. & Kim, K. W. Electrically controlled magnetization in ferromagnet-topological insulator heterostructures. *Phys. Rev. B* **86**, 161406 (2012).
- Stoehr, J. & Siegmann, H. C. *Magnetism: From Fundamentals to Nanoscale Dynamics* (Springer, 2006).
- Xu, S.-Y. *et al.* Hedgehog spin texture and Berry's phase tuning in a magnetic topological insulator. *Nature Phys.* **8**, 616–622 (2012).
- Yokoyama, T., Zang, J. & Nagaosa, N. Theoretical study of the dynamics of magnetization on the topological surface. *Phys. Rev. B* **81**, 241410 (2010).
- Tserkovnyak, Y. & Loss, D. Thin-film magnetization dynamics on the surface of a topological insulator. *Phys. Rev. Lett.* **108**, 187201 (2012).
- Nogueira, F. S. & Eremin, I. Fluctuation-induced magnetization dynamics and criticality at the interface of a topological insulator with a magnetically ordered layer. *Phys. Rev. Lett.* **109**, 237203 (2012).
- Lauter, V., Ambaye, H., Goyette, R., Hal Lee, W.-T. & Parizzi, A. Highlights from the magnetism reflectometer at the SNS. *Physica B* **404**, 2543–2546 (2009).
- Zhu, T. *et al.* The study of perpendicular magnetic anisotropy in CoFeB sandwiched by MgO and tantalum layers using polarized neutron reflectometry. *Appl. Phys. Lett.* **100**, 202406 (2012).
- Korneev, D. A. *et al.* Absorbing sublayers and their influence on the polarizing efficiency of magnetic neutron mirrors. *Nucl. Instrum. Meth. Phys. Res. B* **63**, 328–332 (1992).

Acknowledgements F.K. thanks L. Fu, V. Madhavan, N. Gedik, B. Sinkovic, Y. Wang and H. Lin for discussions. V.L. thanks S. Nagler for discussions, and H. Ambaye, A. Glavic and the Spallation Neutron Source staff for support. The research conducted at ORNL's Spallation Neutron Source was sponsored by the Scientific User Facilities Division, Office of Basic Energy Sciences, and the US Department of Energy. F.K., P.J.-H., and J.S.M. thank the MIT MRSEC through the MRSEC Program of the National Science Foundation under award number DMR-0819762 (upgrade of the molecular beam epitaxy system) for support. J.S.M. thanks the National Science Foundation (DMR-1207469), Office of Naval Research (N00014-13-1-0301) and the STC Center for Integrated Quantum Materials under National Science Foundation grant DMR-1231319 for support, and the thin-film growth and characterization of the materials used. The hetero-structure characterization was supported by the US Department of Energy, Basic Energy Sciences Office, Division of Material Sciences and Engineering under award number DE-SC0006418 (F.K. and P.J.-H.). B.A.A., M.E.J. and D.H. thank the National Science Foundation under award numbers DMR-0907007 and ECCS-1402738 (for SQUID magnetometry characterization) for support. B.A.A. is also supported in part by the Agence Nationale de la Recherche LabEx grants ENS-ICFP (ANR-10-LABX-0010/ANR-10-IDEX-0001-02 PSL). The use of the Advanced Photon Source was supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences, under contract number DE-AC02-06CH11357. I.E. and F.S.N. acknowledge the German Research Council (DFG) for the financial support under the collaborative research centre SFB TR 12 and the priority programme SPP 1666 (grant number ER 463/9).

Author Contributions The research was conceived and designed by F.K. and J.S.M. The samples were prepared and characterized by F.K. The XRD experiments and data analysis were carried out by F.K.; the high-resolution TEM experiments and data analysis were carried out by B.S.; the PNR experiments and data analysis were carried out by V.L.; the XAS/XMCD experiments and data analysis were carried out by F.K. and J.W.F.; the transport experiments and data analysis were carried out by F.K. and D.H.; and the SQUID experiments and data analysis were carried out by F.K., B.A.A., M.E.J. and D.H. The data was interpreted by F.K., V.L., F.S.N. and J.S.M. All authors discussed the results and commented on the manuscript. The manuscript was written by F.K., V.L. and F.S.N.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.K. (katmis@mit.edu) or J.S.M. (moodera@mit.edu).

METHODS

Material growth. The growth of Bi_2Se_3 –EuS bilayer systems were carried out in a molecular beam epitaxy apparatus under an ultrahigh-vacuum environment (10^{-9} – 10^{-10} Torr). High-purity (5 N) Bi and Se constituents were thermally co-evaporated from separate Knudsen cells adjusted to obtain a 2:3 Bi:Se deposition ratio as determined by an *in situ* crystal monitor during growth and confirmed by *ex situ* X-ray reflectivity measurements. (0001)-oriented Al_2O_3 wafers were used as a substrate. To improve the surface quality of the substrates *ex situ* chemical cleaning and *in situ* thermal plus oxygen plasma treatments were performed. After surface preparation, the substrate temperature was kept to $240 \pm 5^\circ\text{C}$ to obtain relatively large surface mobility for epitaxial crystalline Bi_2Se_3 growth. Owing to the high reactivity of Eu atoms and the dissociation problems of S, the EuS was evaporated congruently from a single electron-beam source onto the Bi_2Se_3 layer at a rate of 0.5 – 0.6 \AA s^{-1} without breaking the ultrahigh-vacuum condition. All devices were protected by covering the bilayers with an amorphous Al_2O_3 cap layer via *in situ* electron beam evaporation at room temperature immediately after EuS deposition, without breaking the ultrahigh vacuum.

Interface formation. The interface between EuS and Bi_2Se_3 is analysed using *in situ* reflection high-energy electron diffraction (RHEED). Extended Data Fig. 4a shows a two-dimensional-like (2D-like) (streaky) surface, which is an indication of an atomically flat Bi_2Se_3 surface. EuS is grown at room temperature, which is not enough to give sufficient surface mobility to EuS molecules. Therefore, above a certain critical thickness, which is about 3–4 nm, surface roughening occurs. The RHEED image shows the streaky feature (2D-dominant) for 2-nm-thick EuS grown on Bi_2Se_3 , which we can also call the quasi-2D (2D + 3D) growth mode (Extended Data Fig. 4b and c). After deposition of 5-nm-thick EuS, RHEED images transform from the 2D-dominant to the 3D-dominant phase, which is an indication of a surface roughening (Extended Data Fig. 4d).

X-ray diffraction measurements. A well collimated nearly background-free beam is impinging on the sample surface and X-ray scattering intensity was collected by a two-dimensional charged-coupled device (CCD). The incoming beam is diffracted by a Ge (220) 4-bounce crystal monochromator to obtain $\text{CuK}\alpha_1$ radiation (wavelength $\lambda = 1.54056 \text{ \AA}$). Because of the higher intensity of the Bragg spots, the sample alignments are done on Bi_2Se_3 layer reflections instead of stronger substrate reflections and scans are performed along the Bi_2Se_3 *L*-rod (growth direction). The Bragg reflections are indexed according to the Bi_2Se_3 bulk hexagonal unit cell. The *x*-axis in Fig. 1 is indexed in terms of the hexagonal unit cell of the Bi_2Se_3 , as indicated by $H = K = 0$ with different *L*, where $L = 3, 6, 9, \dots$ are allowed reflections ((0003), (0006), (0009), ...), where *H*, *K*, and *L* are the Miller indices. The Bragg reflection for EuS is calculated by the scattering angles of the peaks and fitted to the bulk EuS unit cell.

TEM. The morphology and structural properties of the layers were separately investigated by scanning TEM and high-resolution TEM. The cross-sectional TEM specimens were prepared using conventional mechanical grinding and dimpling down to below $20 \mu\text{m}$ followed by low-energy (2 keV) and low-angle (4°) Ar-ion milling. TEM images were acquired using a FEI, Tecnai G² F30, S-Twin microscope operating at 300 kV equipped with a Gatan Orius CCD camera.

SQUID magnetometry measurements. The ferromagnetic properties of the Bi_2Se_3 –EuS bilayers were determined by magnetization measurements performed in a Quantum Design SQUID magnetometer. Both in-plane and out-of-plane magnetic properties were measured in the temperature range 2–400 K and applied magnetic fields up to 5 T (Extended Data Fig. 2).

It is known that any distortion reducing the lattice spacing of EuS increases the exchange interaction, thereby increasing both T_C and the spin stiffness²⁷. Our EuS films grown on Bi_2Se_3 have shown compressive stress caused by a $\sim 2\%$ – 10% (depends on the bilayer configuration) lattice mismatch, leading to a reduced lattice spacing. Given the large carrier density of the TI surface present at the TI–FMI interface, this is a possible phenomenon that could enhance the Curie temperature of EuS at the interfacial region. Past studies have shown that the electron doping of Eu chalcogenides can enhance the Curie temperature of the material owing to the increased indirect exchange interaction among Eu^{2+} neighbouring ions^{27–30}. In a recent work, 2% Gd doping was also shown to be effective in increasing the Curie temperature of EuS up to 86.3 K (ref. 31).

The experimental results reported in the main text indicate an extraordinary upwards shift at the interface relative to the bulk value of the Curie temperature in EuS. Typically, all known examples have $\Delta T_C < 1$, whereas the results reported here indicate a shift considerably larger than unity. One example of a large upwards shift in the Curie temperature is provided by $\text{Ni}_3\text{Fe}(111)$, which has a bulk Curie temperature of 850 K, with a surface Curie temperature of 1050 K (ref. 32). The other typical example is the well known Gd(0001) surface, where the bulk and surface Curie temperature values have 293 K and 315 K, respectively³³.

PNR. PNR experiments were performed on the Magnetism Reflectometer at the Spallation Neutron Source at Oak Ridge National Laboratory. Neutrons with wavelengths within a band of 2–8 Å and with a high polarization of 99% to 98.5% were used. Measurements were performed in a closed cycle refrigerator (Advanced Research System CCR) with an applied external magnetic field by using a Bruker electromagnet with a maximum magnetic field of 1.15 T. Using the time-of-flight method²⁴, a collimated polychromatic beam of polarized neutrons with the wavelength band $\Delta\lambda$ impinges on the film at a grazing incidence angle θ , where it interacts with atomic nuclei and the spins of unpaired electrons (see Fig. 3a). The reflected intensity is measured as a function of momentum transfer, $Q = 4\pi\sin(\theta)/\lambda$, for two neutron polarizations R^+ and R^- , with the neutron spin parallel (+) or antiparallel (–) to the direction of the external field, H_{ext} . To separate the nuclear from the magnetic scattering, the data is presented as the spin-asymmetry (SA) ratio $\text{SA} = (R^+(Q) - R^-(Q))/(R^+(Q) + R^-(Q))$ as depicted in Fig. 3d. A value of $\text{SA} = 0$ designates no magnetic moment in the system. Being electrically neutral, spin-polarized neutrons penetrate the entire multilayer structures and probe magnetic and structural composition of the film through the buried interfaces down to the substrate. To show the sensitivity of PNR to the interfacial magnetization measured in 2 QL of Bi_2Se_3 , we intentionally set the magnetization in Bi_2Se_3 to zero (dashed line in the magnetization profile in Fig. 3c) and performed calculations of the corresponding SA (dashed line in Fig. 3d), which shows a considerable deviation from the experimental data in Fig. 3d. PNR results for samples with 10 QL and 5 QL showed similar interfacial magnetization behaviour in Bi_2Se_3 (Extended Data Fig. 3).

To verify the magnetization observed at higher temperatures in the TI–FMI bilayer system, we performed additional measurements with a reference sample of pure EuS film grown on a sapphire substrate under similar conditions. Following the same experimental protocol as for the Bi_2Se_3 –EuS bilayer, the EuS film was cooled at zero magnetic field to 5 K and measured in an external magnetic field of 1 T at different temperature between 5 K and 300 K. Extended Data Fig. 3e represents PNR reflectivity data measured on the EuS film at 5 K, 50 K, 80 K, 120 K, 250 K and 300 K. The data show no difference between R^+ and R^- above 50 K, that is, no magnetization is detected in the pure EuS film above 50 K. In addition, in Extended Data Fig. 3f the experimental data are presented as the SA obtained from the measured reflectivities in Extended Data Fig. 3e. The difference between R^+ and R^- normalized to their sum is very sensitive to small *M* values and serves to emphasize even very small *M*.

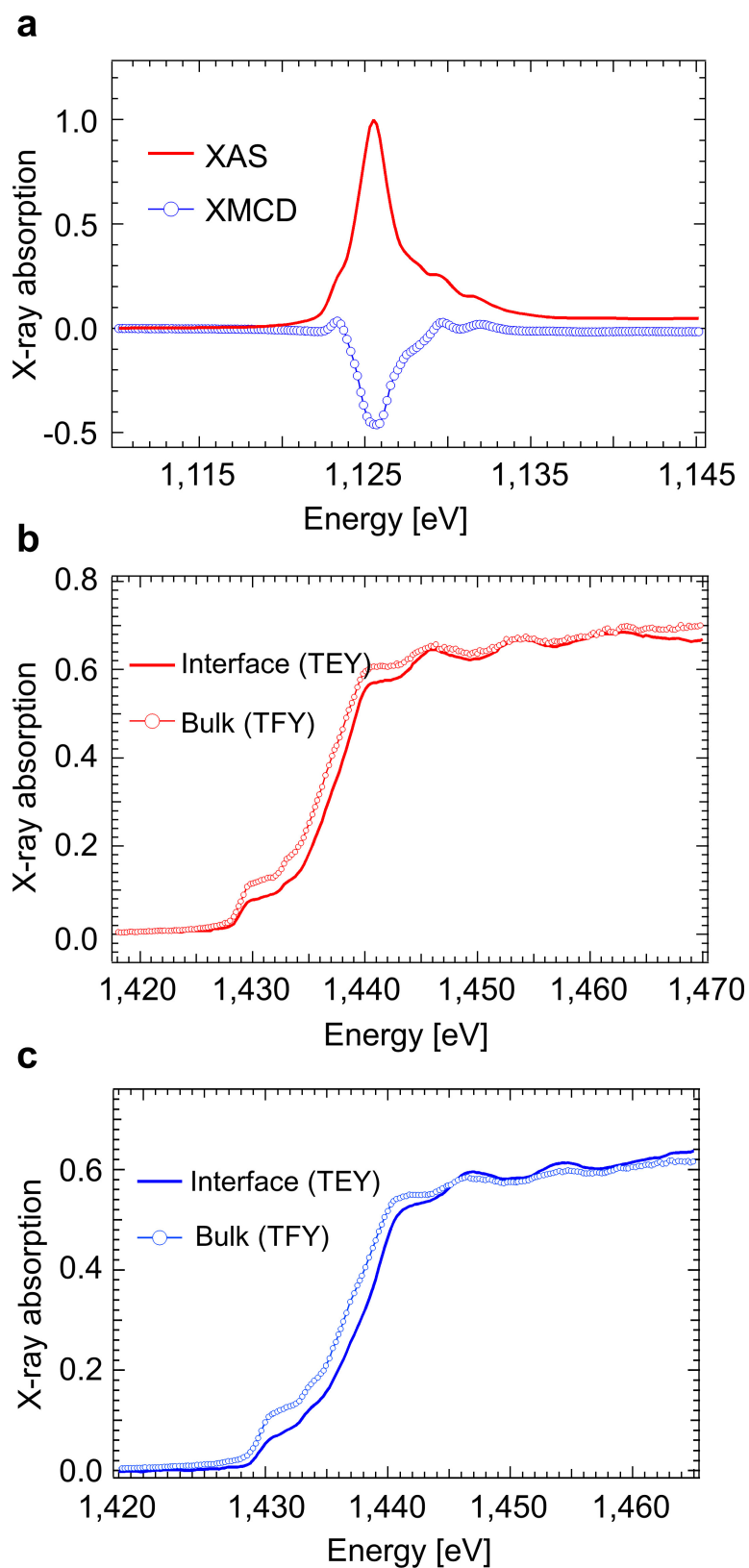
XAS and XMCD. X-ray absorption spectroscopy (XAS) and X-ray magnetic circular dichroism (XMCD) were used to confirm the quality of samples. We performed a series of soft X-ray absorption spectroscopy experiments at beamline 4-ID-C of the Advanced Photon Source by simultaneous measurement of the surface sensitive total electron yield and the bulk sensitive fluorescence yield. First we examine the nature of the magnetic state of the EuS layer. As shown in Extended Data Fig. 1a, the EuS layer is in a valence state of $2+$ (ref. 34) and displays a large XMCD whose lineshape is consistent with a local moment of $7 \mu_B$ per atom³⁵.

To examine the Bi_2Se_3 electronic structure, we measured the Se L_3 edge to probe the electronic structure of the Se. In Extended Data Fig. 1b and c, we compare the bulk-sensitive to the surface/interface-sensitive modes of XAS. For both the $\text{Al}_2\text{O}_3/\text{Bi}_2\text{Se}_3$ (in Extended Data Fig. 1b) and EuS/ Bi_2Se_3 (in Extended Data Fig. 1c) interfaces, the agreement with the bulk is very good, indicating a bulk-like electronic structure on the surface of Bi_2Se_3 . We note that the topological states occur at an energy scale below the resolution of XAS. This data, indicating a sharp electronic interface, agrees well with the diffraction and TEM data, which show the sharp interfacial structure of the bilayer.

Transport measurements. The SQUID magnetometer was also equipped with an electrical probe and used for magnetotransport measurements³⁶. To probe the perpendicular magnetization of the Bi_2Se_3 /EuS interface, bilayer samples were prepared for transport measurements in a Hall bar geometry by mechanically removing areas of the film. Samples were cooled down to 2–5 K in an applied perpendicular field and trained by sweeping the field between ± 5 T. While warming up the samples a 4-T field was applied. Given the small signal, each Hall voltage measurement was averaged ten times at a given field value. Extended Data Fig. 5 displays results from such measurements—the saturating component of the Hall voltage (ΔV_{yx}) for two bilayer samples, 5 QL $\text{Bi}_2\text{Se}_3/5 \text{ nm EuS}$ (Extended Data Fig. 5a–d) and 7 QL $\text{Bi}_2\text{Se}_3/5 \text{ nm EuS}$ (Extended Data Fig. 5e and f), were obtained after subtracting the linear Hall component. Similar trends were observed for the two different samples, showing consistency in the behaviour of the high-temperature ferromagnetic phenomenon. However, it should be mentioned that in Bi_2Se_3 thin films the carrier density is very high ($\sim 10^{18}$ – 10^{19} cm^{-3}), making the bulk contribution to the overall behaviour dominant, especially at higher temperatures where the surface-related effects become masked. (The mobility of the samples increased with film thickness: for example, for 5 QL the mobility was $640 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$, whereas for 20 QL it was $1,650 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$). Hence, precisely

measuring the surface magnetic behaviour by transport is quite challenging, as seen from the Hall data at higher temperatures. In spite of this, further support comes from XMCD measurements, carried out on the same samples, which showed magnetic behaviour similar to the PNR and Hall measurements, and also yielded magnetic moments comparable to those from SQUID measurements. The transport measurements certainly have the resolution to differentiate the interface ferromagnetic behaviour at low temperatures (below ~ 150 K), but not at high temperatures.

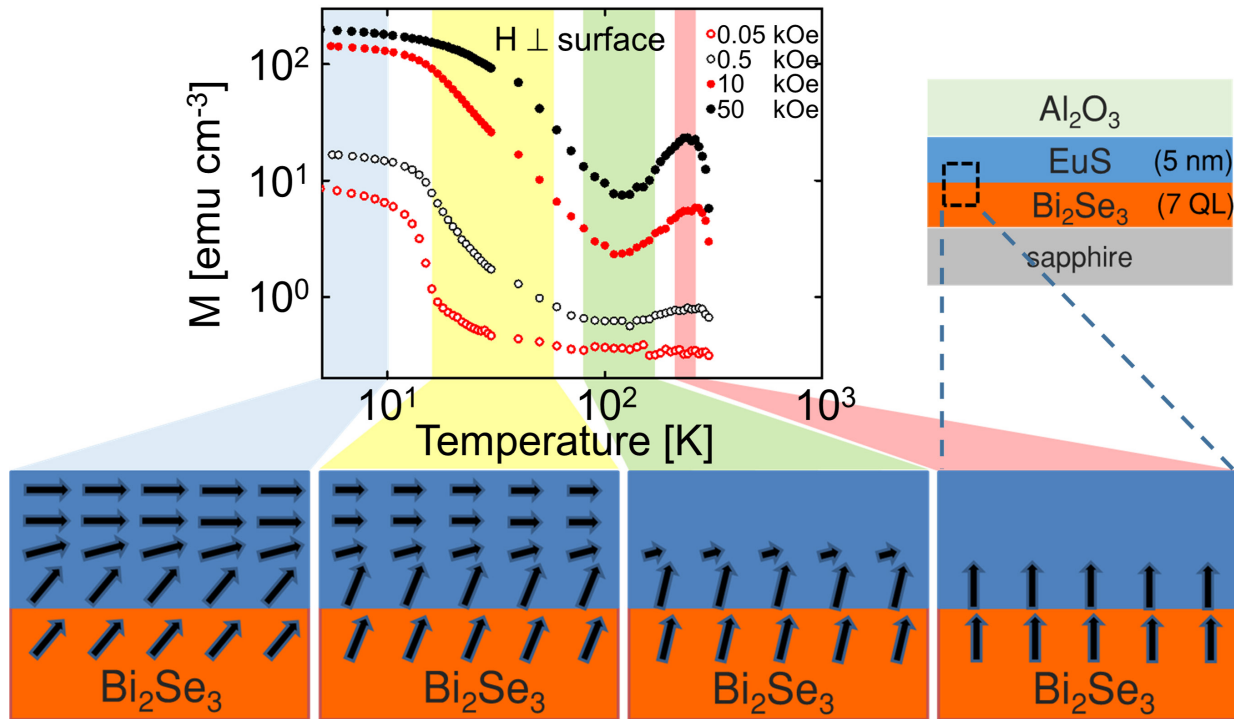
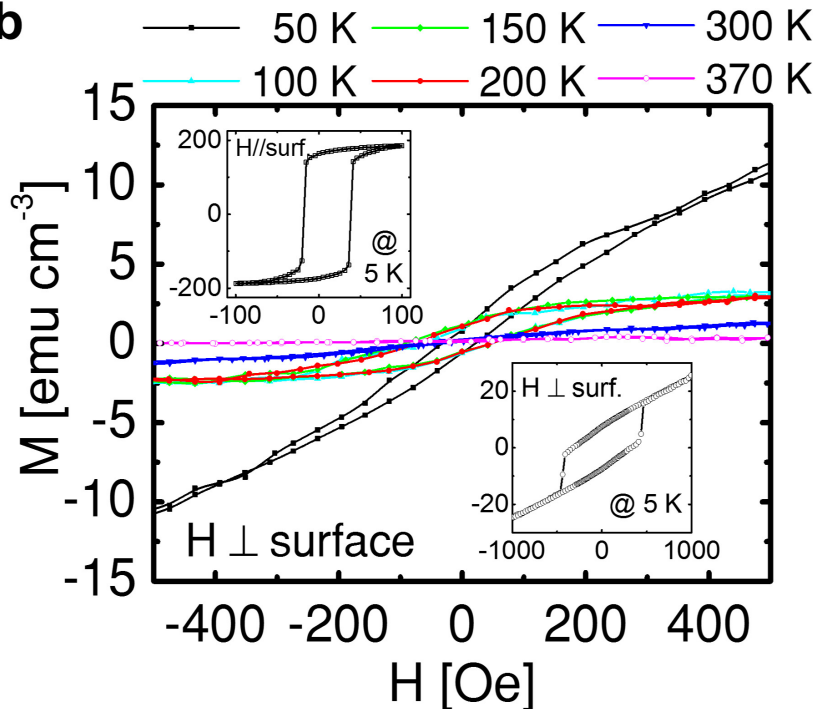
27. Mauger, A. & Godart, C. The magnetic, optical, and transport properties of representatives of a class of magnetic semiconductors: the europium chalcogenides. *Phys. Rep.* **141**, 51–176 (1986).
28. Miyazaki, H. *et al.* La-doped EuO: a rare earth ferromagnetic semiconductor with the highest Curie temperature. *Appl. Phys. Lett.* **96**, 232503 (2010).
29. Ott, H. *et al.* Soft x-ray magnetic circular dichroism study on Gd-doped EuO thin films. *Phys. Rev. B* **73**, 094407 (2006).
30. von Molnár, S. & Kasuya, T. Evidence of band conduction and critical scattering in dilute Eu-chalcogenide alloys. *Phys. Rev. Lett.* **21**, 1757–1761 (1968).
31. Idzuchi, H. *et al.* Critical exponents and domain structures of magnetic semiconductor EuS and Gd-doped EuS films near Curie temperature. *Appl. Phys. Expr.* **7**, 113002 (2014).
32. Mamaev, Y. A., Petrov, V. N. & Starovoitov, S. A. Critical behavior at surfaces. *Sov. Tech. Phys. Lett.* **13**, 642 (1987).
33. Weller, D., Alvarado, S., Gudat, W., Schröder, K. & Campagna, M. Observation of surface-enhanced magnetic order and magnetic surface reconstruction on Gd(0001). *Phys. Rev. Lett.* **54**, 1555–1558 (1985).
34. Kinoshita, T. *et al.* Spectroscopy studies of temperature-induced valence transition on $\text{EuNi}_2(\text{Si}_{1-x}\text{Ge}_x)_2$ around Eu 3d–4f, 4d–4f and Ni 2p–3d excitation regions. *J. Phys. Soc. Jpn.* **71**, 148–155 (2002).
35. Arenholz, E., Schmehl, A., Schlom, D. G. & van der Laan, G. Contribution of Eu 4f states to the magnetic anisotropy of EuO. *J. Appl. Phys.* **105**, 07E101 (2009).
36. Assaf, B. A. *et al.* Modified electrical transport probe design for standard magnetometer. *Rev. Sci. Instr.* **83**, 033904 (2012).



Extended Data Figure 1 | Soft X-ray absorption spectroscopy.

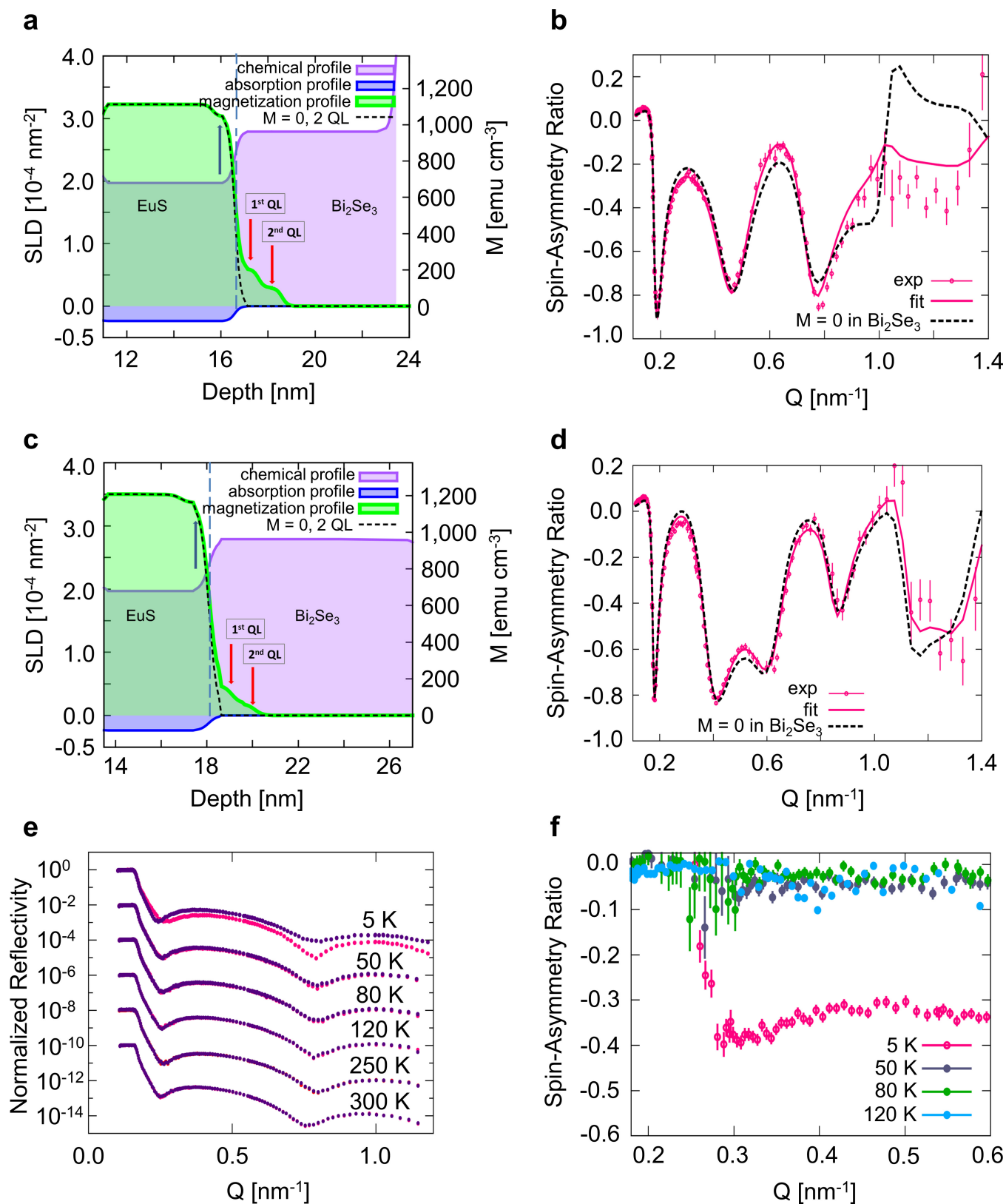
a, Measurement of Eu M_5 edge at 8 K and a field of 2 T. The spectra are consistent with a valence of 2+ and the corresponding large magnetic moment seen by XMCD. **b**, **c**, X-ray absorption at the Se L edge of

$\text{Al}_2\text{O}_3/\text{Bi}_2\text{Se}_3$ (**b**) and $\text{EuS}/\text{Bi}_2\text{Se}_3$ (**c**), showing the good agreement of bulk- and interface-sensitive modes, affirming that the interface and bulk have identical electronic structure. TEY, total electron yield; TFY, total fluorescence yield.

a**b**

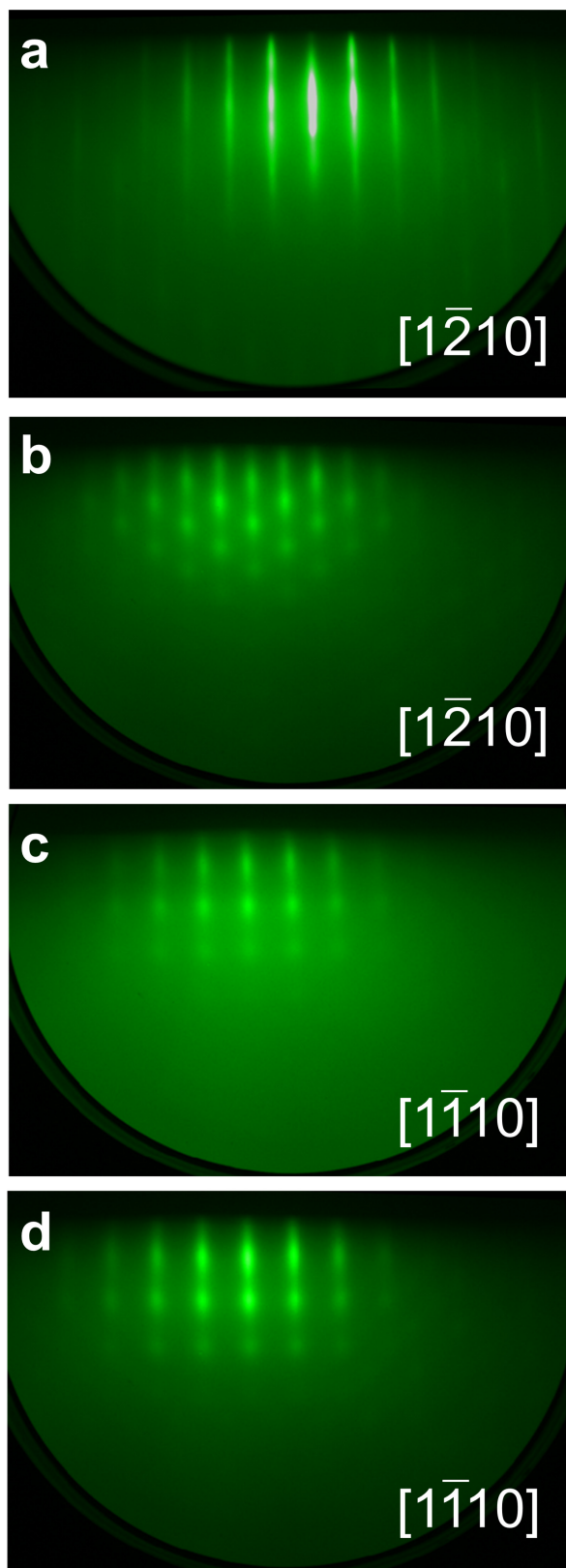
Extended Data Figure 2 | SQUID magnetometry measurements for a Bi_2Se_3 -EuS bilayer with thicknesses of 7 QL for Bi_2Se_3 and 5 nm for EuS. **a**, Magnetization versus temperature at various magnetic fields applied out-of-plane (H perpendicular to the surface). The arrows correspond to the direction of the local magnetization. The large decrease in M as T increases shows the EuS magnetism decreasing (plotted in logarithmic-logarithmic scale). However, at higher temperatures $M(T)$ shows an increase that is much larger than expected from Eu paramagnetism alone, and this could be attributed to reoriented spins (perpendicular) at the interface in the absence of the large in-plane influence from EuS layers above. (Furthermore, control samples of 5-nm-thick EuS grown on

sapphire ($\text{Al}_2\text{O}_3(0001)$) substrate did not show any hysteresis above ~ 50 K even with a 5-T applied field). The possible spin texture is schematically represented below the experimental M versus T results. For the in-plane applied magnetic field configuration, such an increase in magnetization at high temperatures does not show features such as are observed for the perpendicular configuration. The uncertainty in M from the subtraction of the substrate diamagnetism is smaller than the size of the data points. **b**, The low-field magnetic hysteresis at different temperatures, where the field is applied out-of-plane (H perpendicular to the surface). Insets show hysteresis at 5 K comparing data for in-plane (H parallel to the surface) and out-of-plane magnetic-field applications.

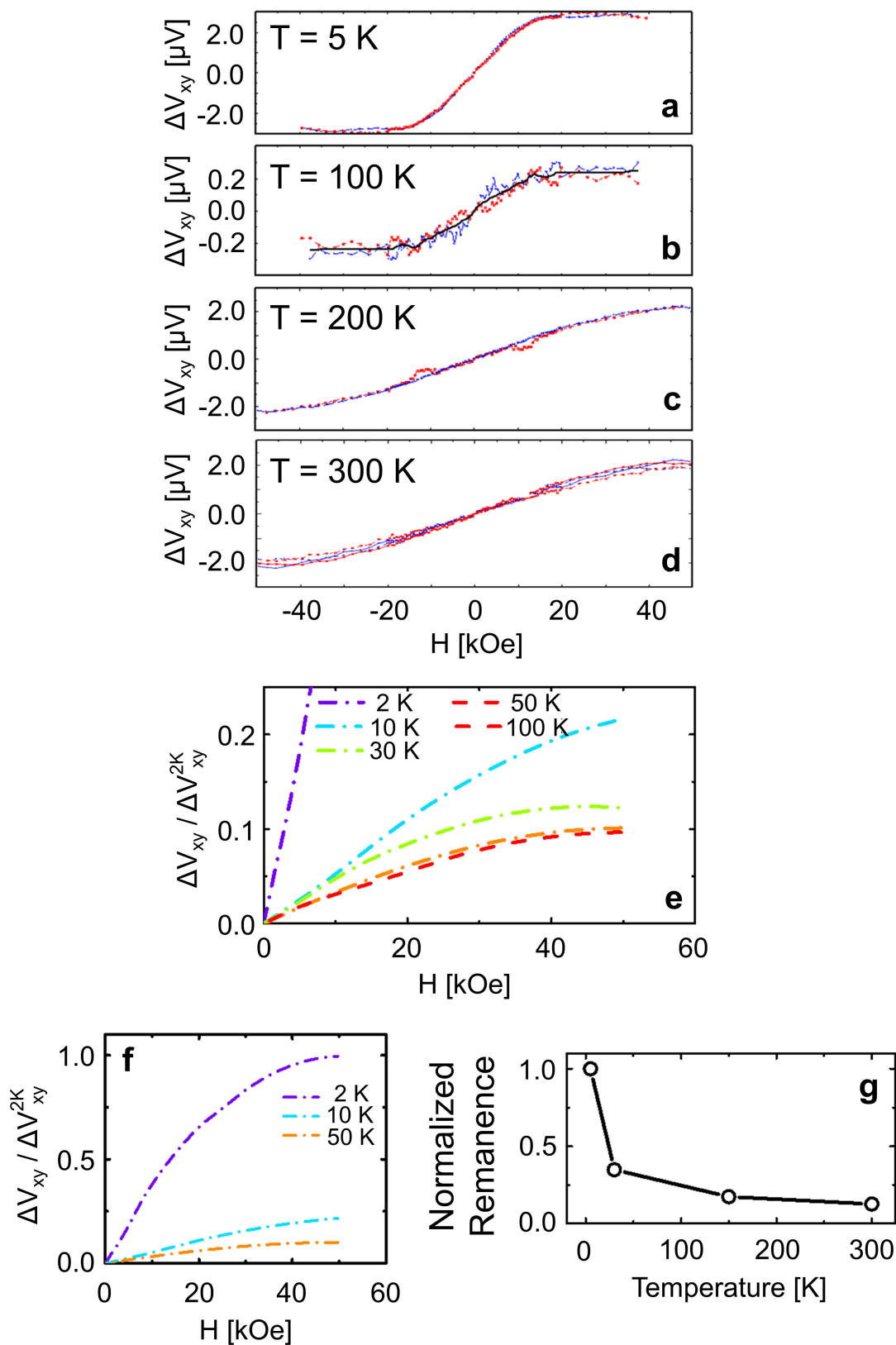


Extended Data Figure 3 | Results from PNR for $\text{Bi}_2\text{Se}_3/\text{EuS}$ bilayer samples with 5 QL, 10 QL of Bi_2Se_3 and pure EuS. a and c, PNR nuclear (NSLD, in pink), magnetic (MSLD, in green) and absorption (ASLD, in blue) scattering length density (SLD) profiles, measured for samples with 5 QL (a) and 10 QL (c) at 5 K and with an in-plane magnetic field of 1 T and presented as a function of the distance from the surface. Magnetization measured inside the Bi_2Se_3 layer is marked with the red arrows. The scale on the right-hand side shows magnetization. **b** and **d**, SA as a function of the momentum transfer Q . Solid curves (dark pink) correspond to the best

fits to the experimental data shown with filled circles with error bars (dark pink), with $\chi^2 = 1.32$ and 1.34, respectively; dashed curves (black) show a considerable deviation from the experimental data when the magnetization in the Bi_2Se_3 2 QL interfacial layer is set to zero with corresponding increased values of $\chi^2 = 2.82$ and 2.56. The error bars represent one standard deviation. **e**, PNR reflectivity data (logarithmic-linear scale) measured on a pure 5-nm-thick EuS film at 5 K, 50 K, 80 K, 120 K, 250 K and 300 K. **f**, Experimental data of the SA obtained from the measured reflectivities in **e**. The error bars represent one standard deviation.



Extended Data Figure 4 | RHEED for interface evolution. **a**, The RHEED pattern for Bi₂Se₃ (2D-like), grown on an Al₂O₃ (0001) surface is shown, where the incident beam is along the [1 $\bar{2}$ 10]-direction of the substrate. The RHEED pattern for the 2-nm-thick EuS surface is shown with the beam along the [1 $\bar{2}$ 10]-direction in **b**, and along the [1 $\bar{1}$ 10]-direction in **c**. The RHEED pattern for the 5-nm-thick EuS surface (3D-dominant) is shown along the [1 $\bar{1}$ 10]-direction in **d**.



Extended Data Figure 5 | Temperature-dependent Hall voltage for a bilayer sample of 7 QL and 5 QL Bi_2Se_3 with 5-nm-thick EuS measured with 10-μA direct current, with magnetic field applied perpendicular to the film plane. A nonlinear contribution to the Hall voltage, ΔV_{xy} , is seen in the 5 QL Bi_2Se_3 /5 nm EuS (a–d) and 7 QL Bi_2Se_3 /5 nm EuS

(e, f) samples. Plot f is the zoom-out of e. The normalized remanent magnetization in the bilayer sample (7 QL Bi_2Se_3 /5 nm EuS) versus temperature (g), shows a finite decrease as temperature increased, matching the Hall data behaviour coming from the interfacial exchange induced ferromagnetic state, as discussed in the main text.

Continuous probing of cold complex molecules with infrared frequency comb spectroscopy

Ben Spaun¹, P. Bryan Changala¹, David Patterson², Bryce J. Bjork¹, Oliver H. Heckl¹, John M. Doyle² & Jun Ye¹

For more than half a century, high-resolution infrared spectroscopy has played a crucial role in probing molecular structure and dynamics. Such studies have so far been largely restricted to relatively small and simple systems, because at room temperature even molecules of modest size already occupy many millions of rotational/vibrational states, yielding highly congested spectra that are difficult to assign. Targeting more complex molecules requires methods that can record broadband infrared spectra (that is, spanning multiple vibrational bands) with both high resolution and high sensitivity. However, infrared spectroscopic techniques have hitherto been limited either by narrow bandwidth and long acquisition time¹, or by low sensitivity and resolution². Cavity-enhanced direct frequency comb spectroscopy (CE-DFCS) combines the inherent broad bandwidth and high resolution of an optical frequency comb with the high detection sensitivity provided by a high-finesse enhancement cavity^{3,4}, but it still suffers from spectral congestion⁵. Here we show that this problem can be overcome by using buffer gas cooling⁶ to produce continuous, cold samples of molecules that are then subjected to CE-DFCS. This integration allows us to acquire a rotationally resolved direct absorption spectrum in the C–H stretching region of nitromethane, a model system that challenges our understanding of large-amplitude vibrational motion^{7–9}. We have also used this technique on several large organic molecules that are of fundamental spectroscopic and astrochemical relevance, including naphthalene¹⁰, adamantane¹¹ and hexamethylenetetramine¹². These findings establish the value of our approach for studying much larger and more complex molecules than have been probed so far, enabling complex molecules and their kinetics to be studied with orders-of-magnitude improvements in efficiency, spectral resolution and specificity.

The massively parallel CE-DFCS technique is virtually equivalent to thousands of simultaneous, highly sensitive, absorption measurements with thousands of narrow linewidth lasers. The broadband (hundreds of nanometres) spectrum of a frequency comb consists of tens of thousands of discrete, narrow frequency modes equally separated by the comb repetition rate, f_{rep} , with a common carrier envelope frequency offset, f_{ceo} . Both f_{rep} and f_{ceo} can be precisely stabilized, allowing for complete control of each of the tens of thousands of separate frequency modes in the comb^{13,14}. By matching evenly spaced comb lines with resonant frequency modes of a high-finesse optical cavity filled with a molecular species, the absorption path length and overall absorption sensitivity of the comb can be enhanced by four orders of magnitude, to kilometre length scales³. A Fourier-transform spectrometer (FTS) is used to measure the absorption spectrum, and only needs a resolution better than the cavity free spectral range (FSR: 100–1,000 MHz) in order to resolve a single comb line. Thus, a standard FTS can achieve an instrument linewidth limited only by the stability of the comb itself, in our case, ~ 50 kHz (ref. 15). The highly multiplexed nature of CE-DFCS has the potential to advance the field of infrared rovibrational spectroscopy just as the recent development of chirped-pulsed

Fourier transform microwave spectroscopy has advanced the field of rotational spectroscopy^{16,17}, but spectral congestion has limited application to relatively simple molecules with less than 10 atoms^{4,5}. This limitation is addressed by combining CE-DFCS with a buffer gas cooling method that can rotationally and translationally cool large molecules to ~ 10 K and below^{18,19}. The simulated nitromethane (CH_3NO_2) absorption spectra in Fig. 1 illustrate the significant gains in resolving power and sensitivity associated with cooling from 300 K to 10 K: molecules have a five-times narrower Doppler-broadened linewidth and occupy many fewer and lower rovibrational energy levels, giving a drastically simplified absorption spectrum (compared to the unresolvable room temperature spectrum²⁰) with clearly distinguishable absorption lines with enhanced peak amplitudes. Supersonic expansion jets^{21,22} provide another means for cooling molecules, but buffer gas cells do not require high pumping speeds and the associated elaborate pumping

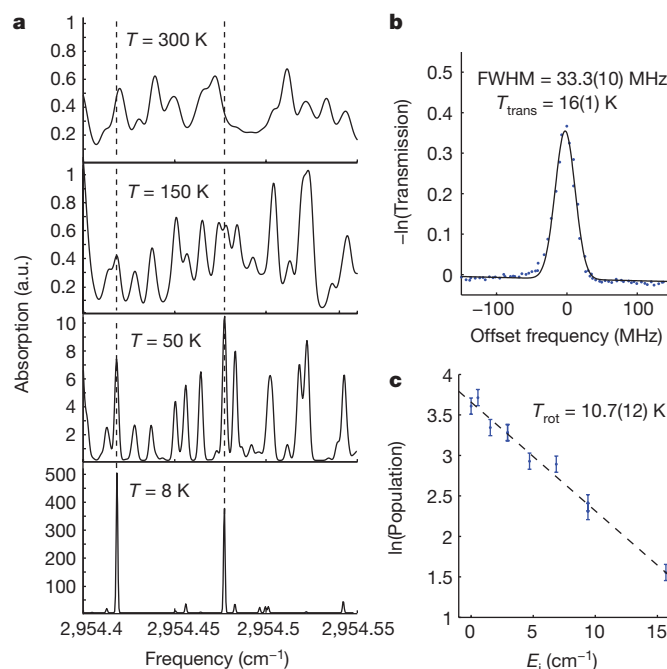


Figure 1 | Buffer gas cooling of nitromethane. **a**, A simulated portion of the nitromethane spectrum as a function of temperature. Individual absorption lines are much more resolvable at low temperatures, as Doppler broadening decreases and the molecular population moves to the lowest available energy levels. **b**, Observed Doppler-broadened absorption profile of nitromethane, showing a translational temperature (T_{trans}) of ~ 16 K. **c**, Measured nitromethane rotational population as a function of energy, E_i , revealing a rotational temperature of $T_{\text{rot}} \approx 11$ K (see Methods for more details). The units cm^{-1} correspond to wavenumber, and are used to represent frequency (**a**) and energy (**c**).

¹JILA, National Institute of Standards and Technology and University of Colorado, Department of Physics, University of Colorado, Boulder, Colorado 80309, USA. ²Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA.

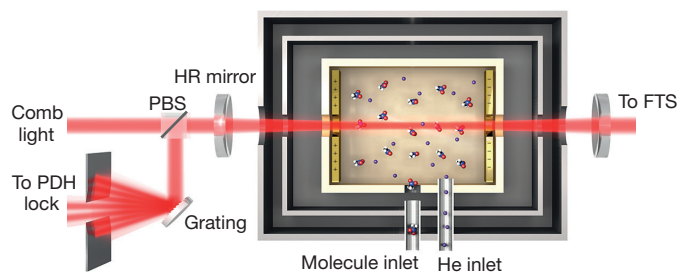


Figure 2 | A schematic of the combined CE-DFCS and buffer gas cooling apparatus. Light from a mid-infrared frequency comb is coupled into a high-finesse enhancement cavity formed by two high reflectivity (HR) mirrors surrounding a 5–10 K buffer gas cell filled with cold molecules. Warm molecules enter through the side of the cell and are quickly cooled to ~ 10 K through multiple collisions with the helium buffer gas, which enters the cell from a separate inlet. Comb light reflected from the cavity is routed by a polarizing beam splitter (PBS) and dispersed by a grating, and an ~ 10 nm segment of the comb spectrum is used to generate a Pound–Drever–Hall (PDH) error signal to lock the comb to the cavity. The PDH lock allows for continuous transmission of thousands of frequency comb modes spanning ~ 100 nm (refs 4, 30). Transmitted comb light is coupled into a Fourier-transform spectrometer (FTS), which measures the fractional absorption of each transmitted comb line.

infrastructure while allowing us to acquire higher-resolution molecular infrared spectra spanning multiple vibrational bands in the mid-infrared with comparable absorption sensitivity.

Figure 2 shows the important components of the combined CE-DFCS and buffer gas cooling apparatus. A mid-infrared frequency comb, tunable from $2.8\text{ }\mu\text{m}$ to $4.8\text{ }\mu\text{m}$, is produced in an optical parametric oscillator (OPO) pumped by a $1\text{ }\mu\text{m}$ ytterbium fibre comb^{4,15}. Both $f_{\text{rep}} \approx 136\text{ MHz}$ and f_{ceo} of the mid-infrared comb are referenced to a microwave caesium clock. The OPO comb output is then coupled into a high finesse ($F \approx 6,000$) optical cavity surrounding a 5–10 K buffer gas cell. The cavity length is servoed via a piezo mirror actuator to ensure that the cavity FSR is always exactly an integer multiple of f_{rep} . This allows many comb frequency modes over a broad bandwidth ($\sim 100\text{ nm}$) to be resonant with the optical cavity⁴. Unlike white light sources, comb light is efficiently coupled into the enhancement cavity because the narrow linewidth of the comb is comparable to that of the cavity. The comb light makes thousands of round trips within the cavity, resulting in a 250 m total absorption path length with cold molecules

in the buffer gas cell. To read out the fractional absorption of each comb mode, we use a custom-built (doubled-passed) fast-scanning FTS with a scanning arm (0.7 m) sufficiently long for single comb mode resolution^{4,5}.

The molecular absorption linewidth ($\Delta\nu$), dominated by 15–30 MHz Doppler broadening (Fig. 1b), is significantly smaller than the frequency spacing between comb modes transmitted through the enhancement cavity (272 MHz). Thus, for given values of f_{rep} and f_{ceo} , the frequency comb is resonant with only a fraction of the molecular absorption features that lie within the comb bandwidth. To ensure that absorption lines are not missed by the discrete frequency comb modes, we step the comb repetition rate by Δf_{rep} after averaging four FTS data acquisitions ($\sim 30\text{ s}$ total). This shifts the frequency of each comb mode by $n\Delta f_{\text{rep}}$, where $n \approx 10^6$ is the comb mode number. We choose Δf_{rep} such that $n\Delta f_{\text{rep}} \leq \Delta\nu/5$, allowing us to measure the Doppler width, and therefore the translational temperature, of molecules in the buffer gas cell in real time. The complete spectrum containing all absorption lines is then generated by interleaving multiple FTS spectra, each corresponding to a different value of f_{rep} .

We demonstrate the simultaneous advantages in resolution, sensitivity and bandwidth in this cold molecule–comb spectroscopy system by gathering rotationally resolved spectra in the C–H stretching region ($\sim 3.3\text{ }\mu\text{m}$) of nitromethane (CH_3NO_2 , a model system for understanding intramolecular vibrational coupling and large amplitude internal motion^{7–9}), adamantane ($\text{C}_{10}\text{H}_{16}$) and hexamethylenetetramine ($\text{C}_6\text{N}_4\text{H}_{12}$, HMT) for the first time. As shown in Fig. 3a, we clearly resolve over 1,000 nitromethane absorption lines spanning multiple vibrational bands, including the entire fundamental C–H stretching region, with an excellent signal-to-noise ratio for less than three hours of data acquisition. The comb bandwidth is sufficiently large to simultaneously gather spectra containing the $\nu_3 + \nu_6$ (symmetric + antisymmetric) N–O stretching combination band and the ν_1 symmetric C–H stretching band (a small section of the ν_1 band is shown magnified in Fig. 3b). The comb was also tuned to a higher centre frequency to acquire the portion of the spectrum covering both components of the ν_{10} asymmetric C–H stretching band.

Making use of existing nitromethane microwave data to provide ground state combination difference frequencies^{23,24}, we assigned transitions for several hundred mid-infrared absorption lines, including those from excited torsional levels (see Methods and Extended Data Tables 1–3 for assignments, line lists and rotational fits). The assigned rovibrational levels reveal interesting intramolecular rovibrational

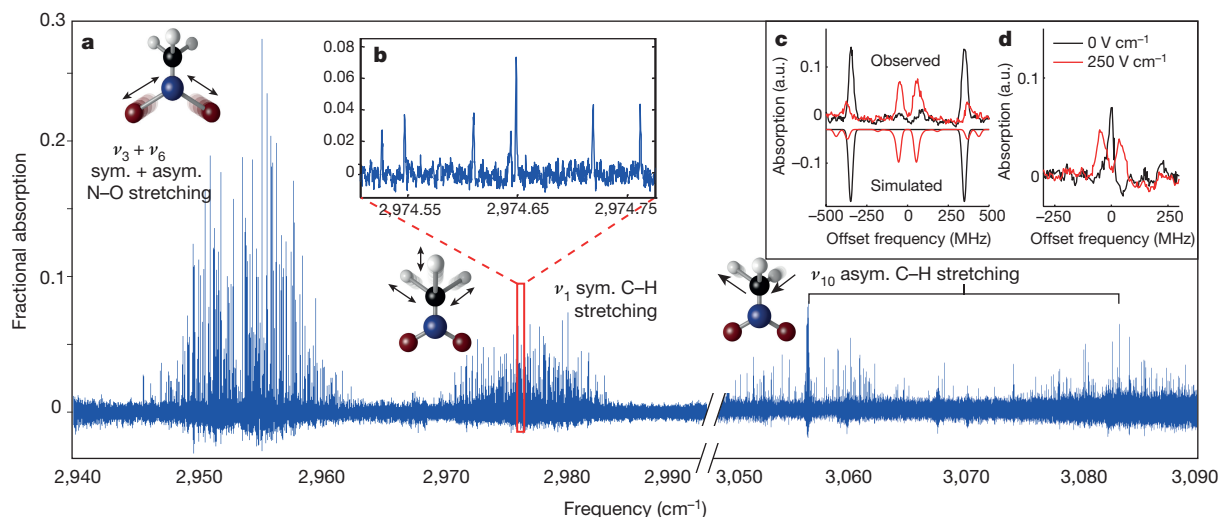


Figure 3 | Survey absorption spectrum of nitromethane. **a**, The spectrum reveals more than 1,000 lines in multiple vibrational bands. The vibrational assignments are indicated, along with a description and illustration of the corresponding vibrational motions associated with each

band (sym., symmetric; asym., asymmetric). **b**, A magnified view of a small section (0.2 cm^{-1}) of the $2,974\text{ cm}^{-1}$ band showing clearly resolved transitions and the typical spectral line density. **c**, **d**, Two examples of characteristic DC Stark splitting patterns, as described in the text.

coupling at play in nitromethane. For example, we observe energy level perturbations characteristic of anharmonic coupling between bright and dark rovibrational states (see Methods and Extended Data Fig. 1 for more detail).

To simplify the line assignment process, we applied a moderate ($50\text{--}400\text{ V cm}^{-1}$) tunable DC electric field within the buffer gas cell and monitored the distinct DC Stark-shift signature of each nitromethane absorption line. An electric field mixes molecular eigenstates together and causes them to experience a unique energy shift that depends on the molecular frame electric dipole moment and the presence of nearby states of opposite parity. As seen in Fig. 3c, d, the high resolution and sensitivity provided by our apparatus allow us to clearly observe these relatively small ($\sim 100\text{ MHz}$) energy level shifts in the nitromethane absorption spectrum. The pattern of the Stark shift is indicative of the specific types of eigenstates participating in the observed molecular transition. The spectrum in Fig. 3d, for example, shows the clear Stark-splitting signature of a transition between excited torsional states ($|m|=1$, where m is the internal rotation quantum number), which split symmetrically and linearly in an electric field^{7,23}. Ground ($m=0$) torsional states, which are non-degenerate, exhibit no such first-order Stark splitting and can therefore be clearly distinguished from excited torsional states. Similarly, closely lying rotational levels of opposite parity will mix together, allowing new transitions to take place between pairs of mixed parity states, an effect clearly observed in the spectrum in Fig. 3c. The apparatus also helped resolve varying degrees of quadratic and linear energy shifts in many other molecular eigenstates when the modest electric field was applied. The information gained by comparing absorption spectra acquired with zero electric field to that acquired with a $50\text{--}400\text{ V cm}^{-1}$ electric field greatly facilitated the assignment process and provided additional confirmation of many line identifications.

Figure 4 shows rotationally resolved spectra of several large organic molecules, including naphthalene (C_{10}H_8), a polyaromatic hydrocarbon with an extensive spectroscopic history^{10,25–27}, adamantane ($\text{C}_{10}\text{H}_{16}$), the simplest member of the diamondoid family¹¹, and HMT ($\text{C}_6\text{N}_4\text{H}_{12}$), a molecule of astrochemical interest^{12,28}. The high-resolution spectra of these molecules, which represent some of the largest molecules to be rotationally resolved in the mid-infrared, were acquired within only 30–90 min per species. To the best of our knowledge, they are the first rotationally resolved spectra for these species in the C–H stretching region^{11,12}, with the exception of naphthalene for which skimmed-molecular-beam optothermal experiments have been performed²⁶, in contrast to the direct absorption spectra we report here. The adamantane spectrum (Fig. 4c) contains over 4,000 absorption features spanning five separate vibrational bands, composed of the three infrared-active C–H stretching fundamentals, ν_{20} , ν_{21} , ν_{22} , and two other bands near $2,853.1\text{ cm}^{-1}$ and $2,904.6\text{ cm}^{-1}$. New, unassigned bands are also observed in naphthalene and HMT near $3,058\text{ cm}^{-1}$ and $2,954\text{ cm}^{-1}$, respectively (Fig. 4a, b). The inset in Fig. 4a shows the reduced term energies measured for rotational levels of the ν_{29} C–H stretching band of naphthalene, illustrating very good agreement with the calculated effective Hamiltonian for a semi-rigid asymmetric top. Furthermore, we have measured the rotational and translational temperatures of these larger molecules and found them to be comparable to those of nitromethane, $\sim 10\text{ K}$. (See Methods for additional analysis of large-molecule spectra and line lists.)

While the cold molecule–comb spectroscopy system enables quick acquisition of rotationally resolved spectra of various large molecules, intramolecular vibrational energy redistribution (IVR) presents an intrinsic challenge to high-resolution spectroscopy of many large molecules in the $3\text{--}5\text{ }\mu\text{m}$ wavelength region of our frequency comb. The low vibrational state density of more rigid molecules, such as naphthalene and adamantane, prevents the onset of severe spectral fractionation due

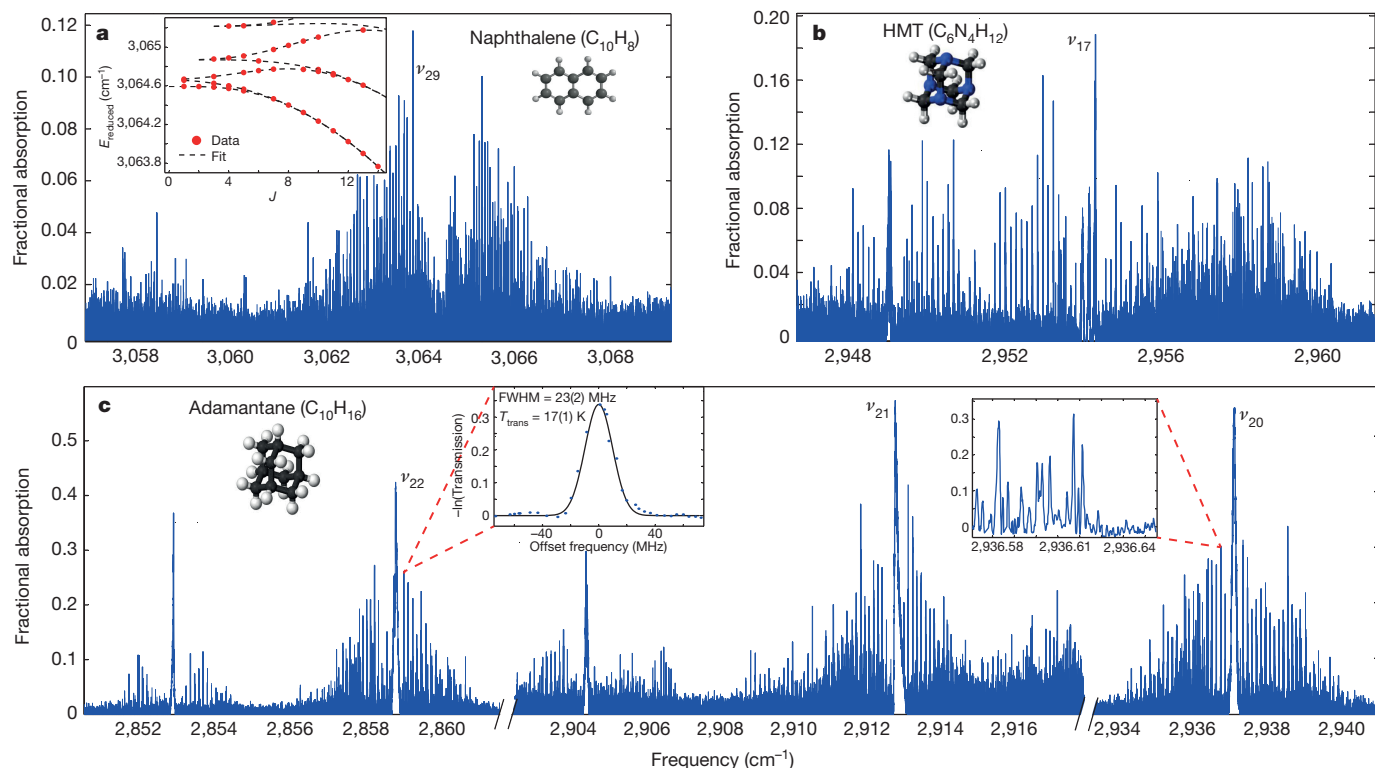


Figure 4 | Survey absorption spectrum of several large molecules. **a–c**, Results for naphthalene (C_{10}H_8 ; **a**), hexamethylenetetramine (HMT, $\text{C}_6\text{H}_{12}\text{N}_4$; **b**) and adamantane ($\text{C}_{10}\text{H}_{16}$; **c**) in the C–H stretching fundamental region. In total, over 4,000 absorption features were resolved in adamantane in 90 min of acquisition time, $\sim 1,500$ lines were resolved in HMT in 30 min, and $\sim 1,000$ lines were resolved in naphthalene in 30 min. Known

vibrational assignments are labelled, and insets in **c** reveal typical line spacing, Doppler-broadened linewidth, and detection noise floor. The translational temperature (T_{trans}) of adamantane was measured to be $17(1)\text{ K}$. As illustrated by the inset of **a**, the observed reduced rotational term energies (E_{reduced}) of the naphthalene ν_{29} band are well described by a semi-rigid asymmetric-top effective Hamiltonian, with a residual scatter of 13 MHz .

to anharmonic and rovibrational coupling between the observed bright states and the dense bath of dark states (Extended Data Fig. 3). But in the case of molecules with significantly higher state densities, such as pyrene or anthracene, IVR is expected to obscure spectra in the C–H stretching region and rotationally resolved spectra of such molecules may only realistically be obtained at lower frequencies²². As we push to acquire high-resolution spectra of even larger molecules in the 3 μm region, highly symmetric, rigid species, such as dodecahedrane, $\text{C}_{20}\text{H}_{20}$, are the most promising targets. Many more molecules, such as C_{60} , will become accessible to this spectroscopic method as frequency comb technology is pushed deeper into the infrared.

We note in closing that while the present data document the successful integration of CE-DFCS and buffer gas cooling to quickly resolve the rovibrational structure of large molecular species, we expect that both the resolving power and detection sensitivity of our system can be significantly improved and its range of application expanded. Regarding the latter, we anticipate that simple modifications to the buffer gas cell will allow us to produce and study cold molecular radicals and that the long (> 10 ms) and continuous interrogation time it provides will open the door to kinetic studies of cold radical reactions using time-resolved frequency comb spectroscopy²⁹.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 September 2015; accepted 15 February 2016.

Published online 4 May 2016.

- Gagliardi, G. & Looch, H.-P. (eds) *Cavity-Enhanced Spectroscopy and Sensing* Chs 4–7 (Springer, 2014).
- Griffith, P. R. & Haseth, J. A. *Fourier Transform Infrared Spectrometry* (Wiley, 2007).
- Thorpe, M. J. *et al.* Broadband cavity ringdown spectroscopy for sensitive and rapid molecular detection. *Science* **311**, 1595–1599 (2006).
- Foltynowicz, A. *et al.* Cavity-enhanced optical frequency comb spectroscopy in the mid-infrared application to trace detection of hydrogen peroxide. *Appl. Phys. B* **110**, 163–175 (2013).
- Adler, F. *et al.* Mid-infrared Fourier transform spectroscopy with a broadband frequency comb. *Opt. Express* **18**, 21861–21872 (2010).
- Patterson, D., Tsikata, E. & Doyle, J. M. Cooling and collisions of large gas phase molecules. *Phys. Chem. Chem. Phys.* **12**, 9736–9741 (2010).
- Tannenbaum, E., Myers, R. J. & Gwinn, W. D. Microwave spectra, dipole moment, and barrier to internal rotation of CH_3NO_2 and CD_3NO_2 . *J. Chem. Phys.* **25**, 42–47 (1956).
- Sørensen, G. O. & Pedersen, T. Symmetry and microwave spectrum of nitromethane. *Stud. Phys. Theor. Chem.* **23**, 219–236 (1983).
- Dawadi, M. B. *et al.* High-resolution Fourier transform infrared synchrotron spectroscopy of the NO_2 in-plane rock band of nitromethane. *J. Mol. Spectrosc.* **315**, 10–15 (2015).
- Albert, S. *et al.* Synchrotron-based highest resolution Fourier transform infrared spectroscopy of naphthalene (C_{10}H_8) and indole ($\text{C}_8\text{H}_7\text{N}$) and its application to astrophysical problems. *Faraday Discuss.* **150**, 71–99 (2011).
- Pirali, O. *et al.* Rotationally resolved infrared spectroscopy of adamantane. *J. Chem. Phys.* **136**, 024310 (2012).
- Pirali, O. & Boudon, V. Synchrotron-based Fourier transform spectra of the ν_{23} and ν_{24} IR bands of hexamethylenetetramine $\text{C}_6\text{H}_{12}\text{N}_4$. *J. Mol. Spectrosc.* **315**, 37–40 (2015).
- Udem, T. *et al.* Absolute optical frequency measurement of the cesium D1 line with a mode-locked laser. *Phys. Rev. Lett.* **82**, 3568–3571 (1999).
- Diddams, S. A. *et al.* Direct link between microwave and optical frequencies with a 300 THz femtosecond laser comb. *Phys. Rev. Lett.* **84**, 5102–5105 (2000).
- Adler, F. *et al.* Phase-stabilized, 1.5 W frequency comb at 2.8–4.8 μm . *Opt. Lett.* **34**, 1330–1332 (2009).
- Brown, G. G. *et al.* A broadband Fourier transform microwave spectrometer based on chirped pulse excitation. *Rev. Sci. Instrum.* **79**, 053103 (2008).
- Park, G. B. *et al.* Design and evaluation of a pulsed-jet chirped-pulse millimeter-wave spectrometer for the 70–102 GHz region. *J. Chem. Phys.* **135**, 024202 (2011).
- Patterson, D. & Doyle, J. M. Cooling molecules in a cell for FTMW spectroscopy. *Mol. Phys.* **110**, 1757–1766 (2012).
- Piskorski, J. *et al.* Cooling, spectroscopy and non-sticking of trans-stilbene and Nile Red. *ChemPhysChem* **15**, 3800–3804 (2014).
- Cavagnat, D. & Lespade, L. Internal dynamics contributions to the CH stretching overtone spectra of gaseous nitromethane NO_2CH_3 . *J. Chem. Phys.* **106**, 7946–7957 (1997).
- Davis, S. *et al.* Jet-cooled molecular radicals in slit supersonic discharges: sub-Doppler infrared studies of methyl radical. *J. Chem. Phys.* **107**, 5661–5675 (1997).
- Brumfield, B. E., Stewart, J. T. & McCall, B. J. Extending the limits of rotationally resolved absorption spectroscopy: pyrene. *J. Phys. Chem. Lett.* **3**, 1985–1988 (2012).
- Rohart, F. Microwave spectrum of nitromethane internal rotation Hamiltonian in the low barrier case. *J. Mol. Spectrosc.* **57**, 301–311 (1975).
- Sørensen, G. O. *et al.* Microwave spectra of nitromethane and D3-nitromethane. *J. Mol. Struct.* **97**, 77–82 (1983).
- Pimentel, G. C. & McClellan, A. L. The infrared spectra of naphthalene crystals, vapor, and solutions. *J. Chem. Phys.* **20**, 270–277 (1952).
- Hewett, K. B. *et al.* High resolution infrared spectroscopy of pyrazine and naphthalene in a molecular beam. *J. Chem. Phys.* **100**, 4077–4086 (1994).
- Pirali, O. *et al.* The far infrared spectrum of naphthalene characterized by high resolution synchrotron FTIR spectroscopy and anharmonic DFT calculations. *Phys. Chem. Chem. Phys.* **15**, 10141–10150 (2013).
- Muñoz Caro, G. M. *et al.* UV-photoprocessing of interstellar ice analogs: detection of hexamethylenetetramine-based species. *Astron. Astrophys.* **413**, 209–216 (2004).
- Fleisher, A. J. *et al.* Mid-infrared time-resolved frequency comb spectroscopy of transient free radicals. *J. Phys. Chem. Lett.* **5**, 2241–2246 (2014).
- Thorpe, M. J. & Ye, J. Cavity-enhanced direct frequency comb spectroscopy. *Appl. Phys. B* **91**, 397–414 (2008).

Acknowledgements We acknowledge funding from DARPA SCOUT, AFOSR, NIST and NSF-JILA PFC for this research. J.M.D. and D.P. acknowledge funding from the NSF and HQOC. B.S. is supported through an NRC Postdoctoral Fellowship. O.H.H. is partially supported through a Humboldt Fellowship. P.B.C. is supported by the NSF GRFP (award no. DGE1144083). We thank J. Baraban for input and discussion. We thank D. Perry for providing us with G. O. Sørensen's original nitromethane ground state data.

Author Contributions P.B.C., D.P., J.M.D. and J.Y. originally designed this experiment. B.S., P.B.C. and J.Y. discussed and implemented the experimental technique, and B.S. and P.B.C. analysed all data. B.S., P.B.C., B.J.B. and O.H.H. operated laboratory equipment. All authors wrote the paper and contributed to technical discussions regarding this work.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.S. (Spaun@jila.colorado.edu) and J.Y. (Ye@jila.colorado.edu).

METHODS

Apparatus. A system of one-inch-thick stainless steel rods and edge-welded bellows stabilizes the cavity length on a macroscopic scale and mechanically isolates the broadband (3.1–3.5 μm) high reflectivity cavity mirrors from the cold cell. The position and angle of each mirror is controlled by a set of three precision screws. One mirror is mounted to a tube-piezo for fine length adjustment with ~ 1 kHz servo bandwidth. The positions of the mirrors, precision mounts, and tube piezo are fixed on a macroscopic length scale by four 1-inch-thick stainless steel rods which are mechanically isolated from the vacuum apparatus.

The length of the high finesse ($F \approx 6,000$) cavity is served so that the cavity free spectral range (FSR) is always exactly twice the mid-infrared comb repetition rate f_{rep} . This allows comb modes spanning a large bandwidth, which is limited by cavity mirror dispersion to ~ 100 nm, to simultaneously stay resonant with the cavity. Phase modulation for the Pound–Drever–Hall (PDH) error signal is obtained by dithering the pump laser cavity length using a fast PZT at one of its resonance frequencies (760 kHz). The light reflected from the cavity is picked off with a polarizing beam splitter, dispersed with a reflection grating, passed through a slit, and directed on a photodiode. The grating and slit serve to select the comb spectral elements to which the cavity is locked. The photodiode signal from the ~ 10 nm wide portion of comb light that passes through the slit is demodulated at the 760 kHz dither frequency to yield an error signal. This error signal is then used to servo the cavity length, via the tube piezo, such that the cavity FSR = $2f_{\text{rep}}$. Following the technique of ref. 15, the mid-infrared comb repetition rate (f_{rep}) and carrier envelope offset frequency (f_{ceo}) are each referenced to a frequency generated by a direct digital synthesizer (DDS) locked to a caesium clock⁴.

Within the buffer gas cell, cold helium gas is used to cool gas-phase molecules to ~ 10 K (ref. 18). The 5–10 K aluminium cell ($6\text{ cm} \times 6\text{ cm} \times 6\text{ cm}$) is anchored to the cold stage of a pulse tube refrigerator and surrounded by a 35 K copper shield to minimize radiative blackbody heating. Electrically insulated flat copper electrodes on opposite sides of the cell allow for the application of tunable DC electric fields parallel to the cavity axis. Between the cell and shield are helium cryopumps made of charcoal. These cryogenic components are enclosed within an $\sim 10^{-6}$ torr vacuum chamber. Warm molecules enter the cell through a small ~ 300 K tube, while a cold (5–10 K) tube delivers the buffer gas. The hot tube must be recessed 1–2 cm from the outer cell wall to prevent parasitic heating of the buffer gas. To achieve sufficiently high inlet flows of larger hydrocarbons, which are solid at room temperature, these species are first vaporized in a 50–200 °C copper oven located just outside the 35 K blackbody shield. When the oven is sufficiently hot, a continuous flow of hydrocarbons exits the oven through a 2 mm aperture and then enters the cold cell. Helium and molecules intermix in the cell where multiple cell–He and He–molecule collisions bring the initially warm molecules into thermal equilibrium with the cold cell. Measured molecular rotational and translational temperatures are typically ~ 10 –15 K (Fig. 1b, c), and molecular and helium densities are estimated to be $\sim 5 \times 10^{12}\text{ cm}^{-3}$ and $\sim 10^{14}\text{ cm}^{-3}$, respectively. To study possible molecule clustering, higher polarizability neon is used as a buffer gas by adding thermal standoffs between the buffer gas cell and the refrigerator cold stage and warming the cell with heating resistors to ~ 20 K.

Fourier transform spectral processing. The path length difference $\Delta\ell$ of our Fourier-transform interferometer is sufficiently long to ensure that the Fourier-transform spectrometer linewidth, which is equal to $(\Delta\ell)^{-1}$, is smaller than the spacing between adjacent frequency comb modes transmitted by the enhancement cavity (that is, the cavity FSR). With the achievement of single comb mode resolution, our tabletop apparatus allows us to obtain broadband absorption spectra with an effective instrument linewidth of ~ 50 kHz (ref. 15), more than two orders of magnitude narrower than the 20 MHz resolution of the best available white-light Fourier-transform infrared spectrometers, which use ~ 10 m translation stages and are primarily available at user facilities^{10,31}.

In order to exploit this drastic improvement in resolution, some post-processing must be applied to the acquired spectrum. The length of the interferogram we collect is typically such that the corresponding spacing between adjacent elements in the frequency domain is about 100 MHz. Since this is not an integer fraction of the absorption cavity FSR (~ 272 MHz), the frequencies of the evenly spaced comb modes and the centre frequencies of the Fourier transform spectrum walk on and off from each other. In order to measure the value of the spectrum at the actual frequencies of the comb modes, we resample the complex-valued spectrum via convolution with the instrument lineshape function (a sinc function). This convolution can be performed efficiently, allowing us to easily and repeatedly resample the spectrum in order to locate the centre frequency and intensity of each comb mode. Similar techniques, employing zero-padding of the interferogram, have been used recently for comb mode resolved Fourier transform spectroscopy by other workers as well³².

Absorption sensitivity. The absorption sensitivity of our comb-based spectrometer for the data presented in this work is $4.4 \times 10^{-8}\text{ cm}^{-1}\text{ Hz}^{-1/2}$ for a single comb

line, for 3,300 resolved comb lines, corresponding to $7.6 \times 10^{-10}\text{ cm}^{-1}\text{ Hz}^{-1/2}$ per spectral element (PSE). We note that this PSE sensitivity is an order of magnitude worse than that previously achieved in ref. 4 at $\sim 3.8\text{ }\mu\text{m}$. We believe there are two factors contributing to our lower PSE sensitivity: increased vibration noise of the cavity enclosing the buffer gas cell; and decreased cavity transmission bandwidth due to higher dispersion of our $\sim 3.3\text{ }\mu\text{m}$ cavity mirrors, compared to dispersion of the $\sim 3.8\text{ }\mu\text{m}$ cavity mirrors used in ref. 4.

Rotational fits. Nitromethane has received considerable spectroscopic attention in both the microwave and infrared regions^{7–9,20,23,24,33–41}, with a general focus placed on understanding the unhindered internal rotation dynamics. A complete list of our assigned $m = 0$ and ± 1 transitions of the $2,953\text{ cm}^{-1}$ band of CH_3NO_2 , which we identify as $\nu_3 + \nu_6$, can be found in Extended Data Tables 2 and 3. (Our vibrational mode labelling convention follows that of table 15–5 in ref. 42.) Upper state term values were calculated using our measured transition frequencies and ground state torsion-rotation energies from previous studies^{23,24}. These energies were used to perform a least-squares fit of the $m = 0$ levels, excluding perturbed states. Unfortunately, the number of $m = \pm 1$ assignments is too small to permit a fit of these levels. Watson's asymmetric-top A-reduced Hamiltonian (I' representation)⁴³ was used as follows:

$$\hat{H} = A\hat{J}_x^2 + \frac{B-C}{2}(\hat{J}^2 - \hat{J}_z^2) + \frac{B-C}{4}(\hat{J}_+^2 + \hat{J}_-^2) - \Delta_f\hat{J}^4 - \Delta_K\hat{J}^2\hat{J}_z^2 - \Delta_K\hat{J}_z^4 \\ - \frac{1}{2}[\delta\hat{J}^2 + \delta_K\hat{J}_z^2 + \hat{J}_+^2 + \hat{J}_-^2]_+$$

See ref. 43 for parameter definitions. Given that the levels measured only extend to $J' \leq 8$, the quartic centrifugal distortion (CD) parameters were not well determined and therefore held fixed to their ground state values. The table shown in Extended Data Table 1a summarizes these results for the $\nu_3 + \nu_6$ band. It also includes the rotational parameters of the $m = 0$ manifolds of the ground state and ν_6 fundamental determined by previous studies^{23,40}.

We observe energy level perturbations in the $\nu_3 + \nu_6$ band characteristic of anharmonic coupling between bright and dark rovibrational states. Most rotational term values of this band fit well to a standard Watson A-reduced Hamiltonian. However, some levels are clearly perturbed by the nearby dark states of a separate vibrational level (see Extended Data Fig. 1). While coupling between the $\nu_3 + \nu_6$ level and higher quanta vibrational levels is possible, we suspect that the perturbing dark state is $\nu_5 + \nu_6$ (ν_5 is the CH_3 umbrella bending mode), the only other two-quanta level expected in the $\sim 2,950\text{ cm}^{-1}$ region. In this case the relatively constant magnitude of the splitting between the mixed eigenstates, $\sim 0.14\text{ cm}^{-1}$, suggests that the coupling between the methyl group vibration ν_5 and the nitro group vibration ν_3 , which could manifest via a quartic k_{3566} term in the anharmonic normal coordinate force field, is relatively weak. This is in stark contrast to the large zeroth-order splitting we observe between the in-plane and out-of-plane components of the nominally degenerate ν_{10} C–H stretching band caused by interactions between the methyl and nitro groups. We also observe Coriolis coupling between the torsional, rotational, and vibrational angular momenta in the $\nu_3 + \nu_6$ and ν_{10} bands. The identification of lines and their spectral patterns is greatly simplified by the lack of systematic fluctuations in line intensities owing to the comb's capability of simultaneous acquisition of spectral features across the vast spectral region.

The shifts of the rotational constants from the ground state to $\nu_3 + \nu_6$ are significantly larger in magnitude than those of the ν_6 fundamental. Indeed, they appear larger than can be accounted for by excitation in ν_3 alone, suggesting significant perturbation of this relatively highly excited level. Unfortunately, the ν_3 fundamental has not yet been analysed at high resolution, and so we cannot make a comparison of the rotational structure of both fundamentals and their combination band. Another way to quantify the change in the rotational structure is through the inertial defect ($\Delta_I = I_c - I_a - I_b$, where $I_{a,b,c}$ are the moments of inertia about the a , b and c axes, respectively), which experiences a large negative shift from $+0.203\text{ u}\text{ }\text{\AA}^2$ in the ground state to $-0.330\text{ u}\text{ }\text{\AA}^2$ in $\nu_3 + \nu_6$. A negative change in the inertial defect is consistent with an increase in the torsional potential barrier⁴¹, as the methyl group becomes locked-in with respect to the plane of the C–NO₂ frame. We note that this is also consistent with our very preliminary analysis of currently assigned $m = \pm 1$ levels. However, further progress on $m > 0$ assignments is necessary before more definitive conclusions can be made. The complete analysis, as well as that of the ν_1 and ν_{10} CH stretch bands, will be reported in a future publication. **Naphthalene.** The ν_{29} band of naphthalene was also treated using the asymmetric-top effective Hamiltonian above. The subset of levels used in the fit ($J' \leq 14$, $K'_d \leq 3$), again, did not permit a determination of the quartic CD parameters, which we held fixed to ground state values from ref. 27. A listing of the 155 b -type transitions we assigned and included in this fit is given in Extended Data Table 4. The fitted molecular constants are summarized in Extended Data Table 1b.

The values of the A , B and C rotational constants of the ν_{29} level are similar to the ground state values. The equilibrium geometry of naphthalene is planar, which is consistent with the small inertial defect of the ground state. However, the defect significantly increases upon excitation of the ν_{29} in-plane C–H stretching mode. Rotational constants for this band have been previously measured in a skimmed molecule beam experiment²⁶, but differ significantly from the values reported here. Indeed, the ground state rotational constants from that study do not agree with our observed ground state combination differences, which were well reproduced using the values listed in table 4 from ref. 27.

Rotational temperature. In order to determine the rotational temperature of the buffer gas cooled nitromethane molecules, we first calculated the relative population in different rotational levels of the torsional-vibrational ground state. To do this, Hamiltonian fit parameters (see above) were used to construct rotational Hamiltonian matrices, which were diagonalized to produce rotational eigenfunctions for both the ground and $\nu_3 + \nu_6$ levels. Transition dipole matrix elements between these rotational eigenfunctions can be calculated using well-known direction cosine matrix elements⁴⁴. The peak intensities of P and R-branch transitions with $J'' = 0-8$ and $K''_a = 0$ (which we estimate to have a measurement uncertainty of 10%) were normalized by the square of the corresponding transition dipole matrix elements to generate the relative populations in each level. The logarithm of these relative populations as a function of energy was then fitted to a first-order polynomial (Fig. 1), the slope of which is equal to $-(kT_{\text{rot}})^{-1}$, where k is the Boltzmann constant and T_{rot} is the effective rotational temperature. Our extracted rotational temperature of 10.7(12) K is comparable to the translational temperature, 16(1) K, determined by the measured Doppler widths. Thermalization between translational and rotational degrees of freedom is only partial, but certainly more complete than that typically obtainable in a supersonic jet.

Sample size. No statistical methods were used to predetermine sample size.

Molecule–buffer gas clustering. Because of the low buffer gas temperature, the formation of weakly bound complexes containing a molecule and buffer gas atom(s) is possible. We attempted to observe neon–acetylene, Ne–C₂H₂, complexes by cooling C₂H₂ in a neon buffer gas at 20 K. In Extended Data Fig. 2, we compare the previously measured spectrum of the complex (upper trace, reproduced from ref. 45), with our measured spectrum of the buffer gas cell (lower trace). We observe no absorption from Ne–C₂H₂ above our baseline noise floor. The acetylene flow rate into the cell for this measurement (10 sccm) was sufficiently high to saturate our absorption dynamic range for most of the monomer transitions. Therefore, to aid in the comparison of the relative absorption of the complex, we have marked the R(0) transition at 3,286.476 cm^{−1} of the ν_3 band of HC¹³CH, which occurs at about 1% natural abundance relative to the normal isotopologue. Two hot band transitions, at 3,285.891 and 3,286.176 cm^{−1}, from vibrational levels with one quantum of excitation in either of the two degenerate bending modes ($(\nu_2 + 2\nu_4 + \nu_5)_{\text{II}}^{\ell=1} - \nu_4$, R(6f) and $(\nu_2 + \nu_4 + 2\nu_5)_{\text{II}}^{\ell=1} - \nu_5$, R(5f)) are also labelled in the cold cell spectrum. Based on this measurement, we estimate the peak absorption of Ne–C₂H₂ to be less than 0.1% relative to the monomer. This upper bound complements previous experiments that determined the population fraction of He–*trans*-stilbene complexes to be less than 5% using ultraviolet laser induced fluorescence (UV-LIF) spectroscopy in a similar cold cell apparatus¹⁹.

Vibrational density of states. The vibrational density of states estimates presented in Extended Data Fig. 3 were calculated using a direct state counting algorithm⁴⁶. We used observed—and when not available, calculated—vibrational frequencies for adamantane^{47,48}, naphthalene^{26,49,50}, diamantane^{51,52}, dodecahedrane^{53,54}, anthracene^{55,56}, and pyrene^{57,58}. In the case of only purely vibrational anharmonic interactions, the relevant density of states is that of states with the same vibrational symmetry as the zero-order bright state. This fraction is n^2/g , where g is the order of the molecular point group and n is the dimension of the irreducible representation of interest⁵⁹. For the infrared active C–H stretching fundamental levels, in particular, these fractions are 1/8 for naphthalene, anthracene, and pyrene; 3/8 for adamantane; 3/40 for dodecahedrane; and 1/12 or 1/3 for diamantane (non-degenerate or doubly degenerate modes, respectively).

31. Brubach, J. *et al.* Performance of the AILES THz-infrared beamline at SOLEIL for high resolution spectroscopy. *AIP Conf. Proc.* **1214**, 81–84 (2010).

32. Maslowski, P. *et al.* Surpassing the path-limited resolution of Fourier-transform spectrometry with frequency combs. *Phys. Rev. A* **93**, 021802(R) (2016).

33. Cox, A. P. & Waring, S. Microwave spectrum and structure of nitromethane. *J. Chem. Soc. Faraday Trans. 2* **68**, 1060–1071 (1972).

34. Jones, W. J. & Sheppard, N. The gas-phase infrared spectra of nitromethane and methyl boron difluoride; fine structure caused by internal rotation. *Proc. R. Soc. Lond. A* **304**, 135–155 (1968).

35. McKean, D. C. & Watt, R. A. Vibrational spectra of nitromethanes and the effects of internal rotation. *J. Mol. Spectrosc.* **61**, 184–202 (1976).

36. Hill, J. R. *et al.* Infrared, Raman, and coherent anti-Stokes Raman spectroscopy of the hydrogen/deuterium isotopomers of nitromethane. *J. Phys. Chem.* **95**, 3037–3044 (1991).

37. Gorse, D. *et al.* Theoretical and spectroscopic study of asymmetric methyl rotor dynamics in gaseous partially deuterated nitromethanes. *J. Phys. Chem.* **97**, 4262–4269 (1993).

38. Hazra, A., Ghosh, P. & Kshirsagar, R. Fourier transform infrared spectrum and rotational structure of the A-type 917.5 cm^{−1} band of nitromethane. *J. Mol. Spectrosc.* **164**, 20–26 (1994).

39. Hazra, A. & Ghosh, P. Assignment of the $m = 0$ transitions in the ν_4 band of nitromethane by the symmetric top approximation method. *J. Mol. Spectrosc.* **173**, 300–302 (1995).

40. Pal, C. *et al.* High resolution Fourier transform infrared spectrum and vibration-rotation analysis of the B-type 1584 cm^{−1} band of nitromethane. *J. Mol. Struct.* **407**, 165–170 (1997).

41. Halonen, M. *et al.* Molecular beam infrared spectrum of nitromethane in the region of the first C–H stretching overtone. *J. Phys. Chem. A* **102**, 9124–9128 (1998).

42. Bunker, P. R. & Jensen, P. *Molecular Symmetry and Spectroscopy* 2nd edn (NRC Research Press, Ottawa, 1998).

43. Watson, J. K. G. *Vibrational Spectra and Structure* Vol. 6, Ch. 1 (ed. Durig, J.) (Elsevier, 1977).

44. Townes, C. H. & Schawlow, A. L. *Microwave Spectroscopy* (Dover, 1975).

45. Bemish, R. J. *et al.* Infrared spectroscopy and *ab initio* potential energy surface for Ne–C₂H₂ and Ne–C₂HD complexes. *J. Chem. Phys.* **109**, 8968–8978 (1998).

46. Baer, T. & Hase, W. L. *Unimolecular Reaction Dynamics* (Oxford Univ. Press, 1996).

47. Bistričić, L., Baranović, G. & Mlinarić-Majerski, K. A vibrational assignment of adamantane and some of its isotopomers. Empirical versus scaled semiempirical force field. *Spectrochim. Acta A* **51**, 1643–1664 (1995).

48. Jensen, J. O. Vibrational frequencies and structural determination of adamantane. *Spectrochim. Acta A* **60**, 1895–1905 (2004).

49. Sellers, H., Pulay, P. & Boggs, J. E. Theoretical prediction of vibrational spectra. 2. Force field, spectroscopically refined geometry, and reassignment of the vibrational spectrum of naphthalene. *J. Am. Chem. Soc.* **107**, 6487–6494 (1985).

50. Mitra, S. S. & Bernstein, H. J. Vibrational spectra of naphthalene-d₀, - α -d₄, and -d₈ molecules. *Can. J. Chem.* **37**, 553–562 (1959).

51. Ramachandran, G. & Manogaran, S. Vibrational spectra of adamantanes X₁₀H₁₆ and diamantanes X₁₄H₂₀ (X = C, Si, Ge, Sn): a theoretical study. *J. Mol. Struct. THEOCHEM* **766**, 125–135 (2006).

52. Jenkins, T. & Lewis, J. A Raman study of adamantane (C₁₀H₁₆), diamantane (C₁₄H₂₀) and triamantane (C₁₈H₂₄) between 10 K and room temperatures. *Spectrochim. Acta A* **36**, 259–264 (1980).

53. Hudson, B. S. *et al.* Infrared, Raman, and inelastic neutron scattering spectra of dodecahedrane: an *I_h* molecule in *T_h* site symmetry. *J. Phys. Chem. A* **109**, 3418–3424 (2005).

54. Karpushenkova, L. S., Kabo, G. J. & Bazyleva, A. B. Structure, frequencies of normal vibrations, thermodynamic properties, and strain energies of the cage hydrocarbons C_nH_n in the ideal-gas state. *J. Mol. Struct. THEOCHEM* **913**, 43–49 (2009).

55. Szczepanski, J. *et al.* Electronic and vibrational spectra of matrix isolated anthracene radical cations: experimental and theoretical aspects. *J. Chem. Phys.* **98**, 4494–4511 (1993).

56. Bakke, A. *et al.* Condensed aromatics. Part II. The five-parameter approximation of the in-plane force field of molecular vibrations. *Z. Naturforsch. C* **34a**, 579–584 (1979).

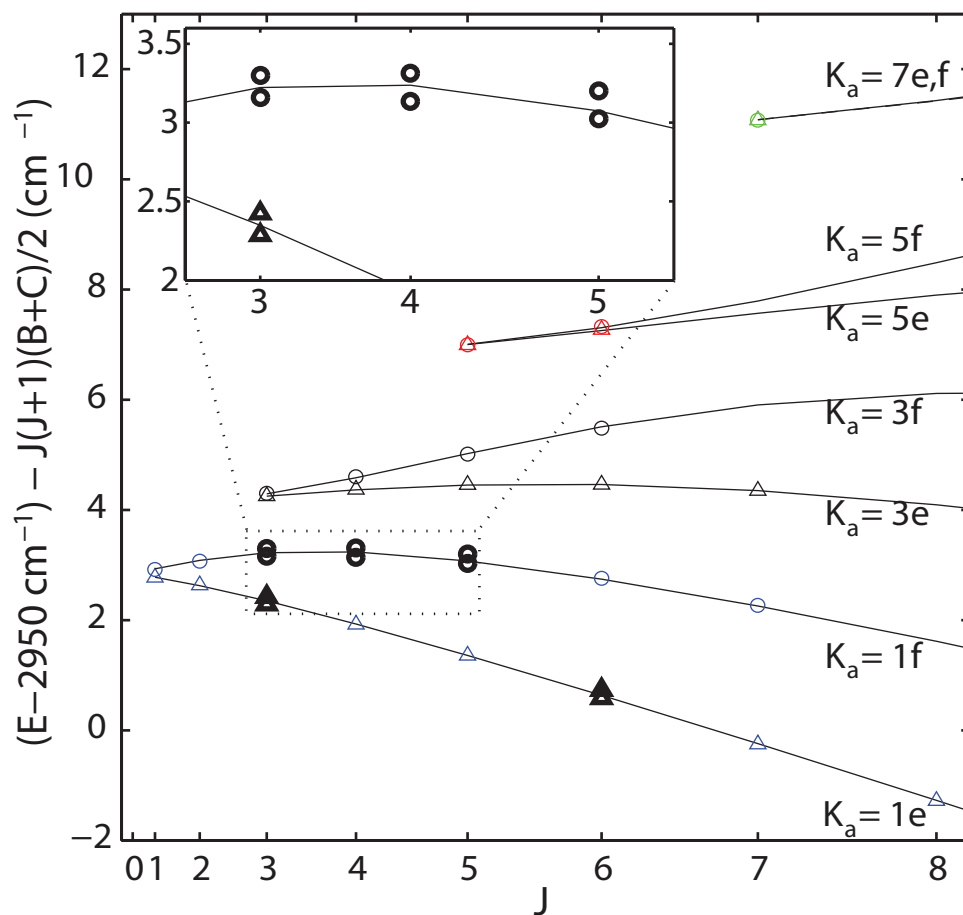
57. Vala, M. *et al.* Electronic and vibrational spectra of matrix-isolated pyrene radical cations: theoretical and experimental aspects. *J. Phys. Chem.* **98**, 9187–9196 (1994).

58. Shinohara, H., Yamakita, Y. & Ohno, K. Raman spectra of polycyclic aromatic hydrocarbons. Comparison of calculated Raman intensity distributions with observed spectra for naphthalene, anthracene, pyrene, and perylene. *J. Mol. Struct.* **442**, 221–234 (1998).

59. Pechukas, P. Comment on “Densities of vibrational states of given symmetry species”. *J. Phys. Chem.* **88**, 828 (1984).

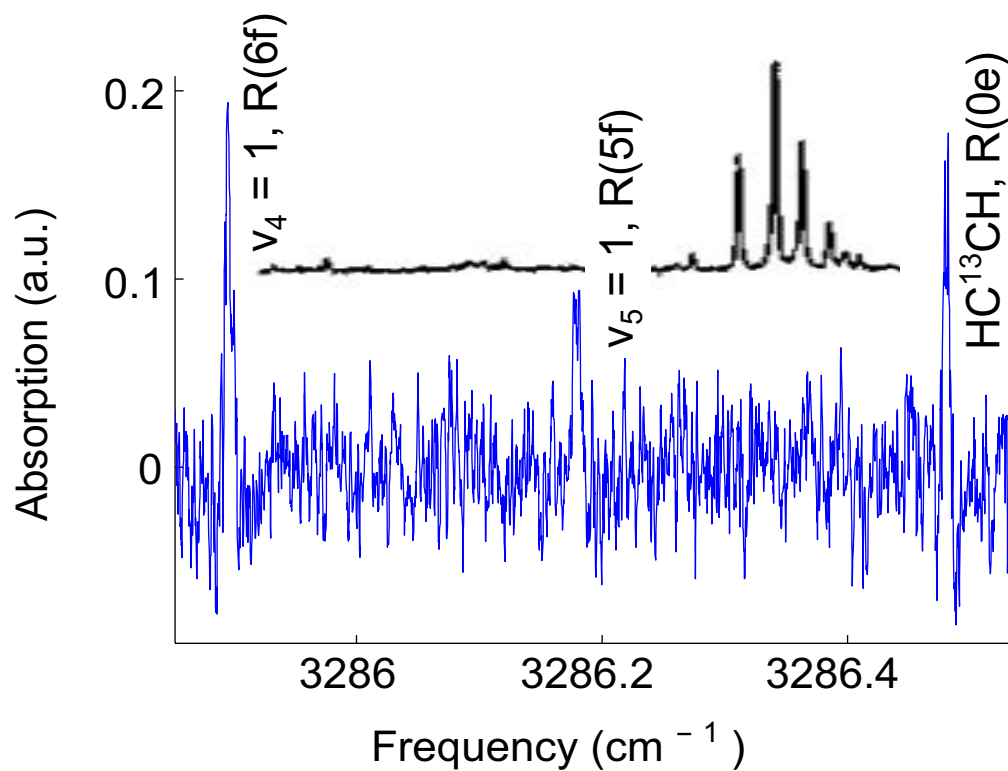
60. Nesbitt, D. J. & Field, R. W. Vibrational energy flow in highly excited molecules: role of intramolecular vibrational redistribution. *J. Phys. Chem.* **100**, 12735–12756 (1996).

61. Buckingham, G. T., Chang, C.-H. & Nesbitt, D. J. High-resolution rovibrational spectroscopy of jet-cooled phenyl radical: the ν_{19} out-of-phase symmetric CH stretch. *J. Phys. Chem. A* **117**, 10047–10057 (2013).



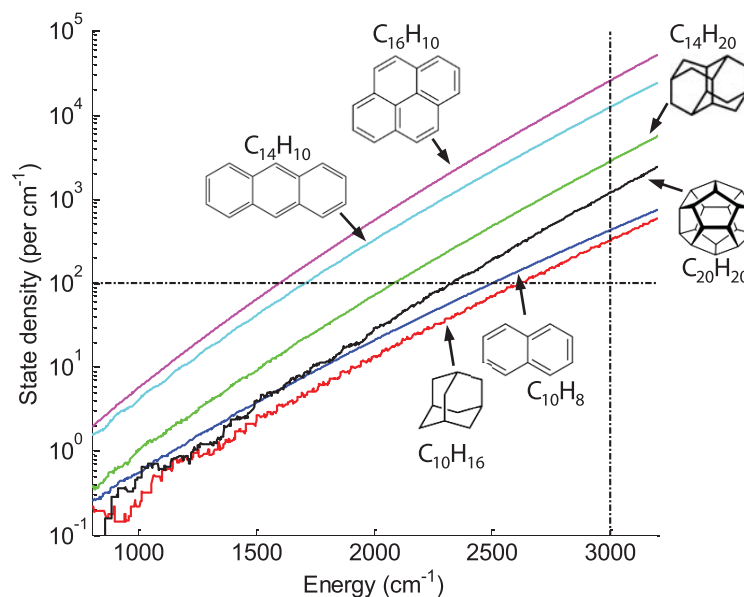
Extended Data Figure 1 | Reduced term values of the rotational sub-levels of $\nu_3 + \nu_6$ ($m = 0$). These are plotted against the total angular momentum, J (scaled as $J(J+1)$). The reduced energies are equal to the absolute energy E , offset by $2,950 \text{ cm}^{-1}$, minus $J(J+1)$ times the average of the B and C rotational constants. The solid lines connect sets of levels with respect to K_a (the projection of J onto the molecular inertial a axis)

and their parity (e/f) symmetry label. For clarity, e and f states are shown in triangles and circles, respectively. States of different K_a values are shown in different colours. Inset, magnified view of the boxed area of the main plot, showing pairs of perturbed eigenstates, split symmetrically about the zeroth-order bright state position, are indicated in bold markers (see Methods for additional details).



Extended Data Figure 2 | Evidence of cluster-free cooling. The plot compares our measured buffer gas cooled C_2H_2 spectrum (bottom trace) with that of the $\text{Ne-C}_2\text{H}_2$ complex (upper trace; reprinted with permission from figure 1 of ref. 45, copyright 1998, AIP Publishing LLC).

Three acetylene monomer transitions in the buffer gas cooled spectrum, including two hot band transitions and a ^{13}C feature as described in the text, have been labelled. The buffer gas cooled spectrum has been rebinned with a bin size of 5 frequency elements (~ 40 MHz total).



Extended Data Figure 3 | The vibrational density of states for several large hydrocarbons. In increasing order, the total density of states (that is, not symmetry selected) versus vibrational energy is shown for adamantane ($C_{10}H_{16}$), naphthalene ($C_{10}H_8$), dodecahedrane ($C_{20}H_{20}$), diamantane ($C_{14}H_{20}$), anthracene ($C_{14}H_{10}$), and pyrene ($C_{16}H_{10}$). These curves were calculated using a direct state count algorithm and a combination of

previously observed and calculated vibrational frequencies (see Methods for details). The horizontal line at 100 states per cm^{-1} marks the empirical threshold symmetry selected state density for IVR^{60,61}. The vertical line at 3,000 cm^{-1} indicates the approximate energy for CH stretch fundamental vibrations.

Extended Data Table 1 | Rotational Hamiltonian fit results

a	Parameter	Ground state *	$v_3 + v_6^\dagger$	v_6^\ddagger
	T_v	0	2952.6854(45)	1583.81163(20)
	A	0.44503725(100)	0.439902(180)	0.4449620(33)
	B	0.35172249(100)	0.347716(303)	0.3516825(26)
	C	0.19599426(97)	0.194949(143)	0.1960255(9)
	$\Delta_J \times 10^6$	0.2048(207)	[0.2048]	0.2431(23)
	$\Delta_{JK} \times 10^6$	0.5921(123)	[0.5921]	0.6822(103)
	$\Delta_K \times 10^6$	-0.2515(133)	[-0.2515]	-1.5701(93)
	$\delta_J \times 10^6$	0.08229(147)	[0.08229]	0.0717(11)
	$\delta_K \times 10^6$	0.52536(634)	[0.52536]	0.4573(34)
	RMSE	—	1.1×10^{-2}	2.3×10^{-3}
	Δ_i	+0.203	− 0.330	+0.177

b	Parameter	Ground state §	v_{29}^{\parallel}	v_{29}^{\P}
	T_v	0	3064.5942(5)	3064.58(2)
	A	0.104051836(124)	0.104198(30)	0.104013(17)
	B	0.04112733(37)	0.0411173(38)	0.0411023(45)
	C	0.029483552(140)	0.02942455(9)	0.0294062(20)
	$\Delta_J \times 10^9$	0.528(49)	[0.528]	
	$\Delta_{JK} \times 10^9$	1.206(145)	[1.206]	
	$\Delta_K \times 10^9$	5.648(112)	[5.648]	
	$\delta_J \times 10^9$	0.1752 [#]	[0.1752]	
	$\delta_K \times 10^9$	1.951 [#]	[1.951]	
	RMSE	3.1×10^{-4}	4.3×10^{-4}	6.3×10^{-4}
	Δ_i	− 0.137	+1.137	+1.057

a, Asymmetric-top Hamiltonian fit results for CH_3NO_2 . Values in brackets are fixed, and uncertainties are given in parentheses. All parameter values are specified in cm^{-1} . The inertial defect Δ_i , determined from $I_c - I_a - I_b$, is given in u Å^2 . **b**, Asymmetric-top Hamiltonian fit results for naphthalene. Values in brackets are fixed, and uncertainties are given in parentheses. (See Methods for parameter definitions.)

*Ref. 24.

†This work.

‡Ref. 40.

§Ref. 28.

||This work.

¶Ref. 27.

#Calculated values²⁸.

Extended Data Table 2 | Line list for the 2,953 cm⁻¹ band of nitromethane

m''	J''	Ka''	Kc''	E'' ^a	E'' ^b	m'	J'	Ka'	Kc'	Line Pos.	E' ^a	E' ^b	Comment
0	2	2	0	2.4213	2.421311	0	1	1	1	0.8969	3.3182	3.3182	
0	2	0	2	1.5497	1.549651	0	1	1	1	1.7690	3.3187	3.3186	
0	0	0	0	0.0000	0.000000	0	1	1	1	3.3184	3.3184	3.3184	
0	2	2	1	2.3278	2.327828	0	1	1	0	1.1362	3.4640	3.4640	
0	1	0	1	0.5477	0.547713	0	1	1	0	2.9161	3.4638	3.4638	
0	3	2	1	4.3221	4.322080	0	2	1	2	-0.0584	4.2637	4.2637	
0	3	0	3	2.9351	2.935110	0	2	1	2	1.3287	4.2638	4.2638	
0	2	2	1	2.3278	2.327828	0	2	1	2	1.9358	4.2636	4.2637	
0	1	0	1	0.5477	0.547713	0	2	1	2	3.7159	4.2636	4.2636	
0	3	2	2	3.9710	3.970934	0	2	1	1	0.7220	4.6930	4.6929	
0	2	2	0	2.4213	2.421311	0	2	1	1	2.2719	4.6932	4.6932	
0	2	0	2	1.5497	1.549651	0	2	1	1	3.1437	4.6934	4.6934	
0	4	0	4	4.7032	4.703276	0	3	1	3	0.8403	5.5435	5.5436	component 1
0	3	2	2	3.9710	3.970934	0	3	1	3	1.5730	5.5440	5.5440	
0	2	0	2	1.5497	1.549651	0	3	1	3	3.9944	5.5441	5.5440	
0	4	0	4	4.7032	4.703276	0	3	1	3	0.9769	5.6801	5.6802	component 2
0	3	2	2	3.9710	3.970934	0	3	1	3	1.7096	5.6806	5.6805	
0	2	0	2	1.5497	1.549651	0	3	1	3	4.1310	5.6807	5.6806	
0	4	2	3	6.0822	6.082186	0	3	1	2	0.3326	6.4148	6.4148	component 1
0	3	2	1	4.3221	4.322080	0	3	1	2	2.0930	6.4151	6.4151	
0	3	0	3	2.9351	2.935110	0	3	1	2	3.4799	6.4150	6.4150	
0	2	2	1	2.3278	2.327828	0	3	1	2	4.0869	6.4147	6.4147	
0	4	2	3	6.0822	6.082186	0	3	1	2	0.4721	6.5543	6.5543	component 2
0	3	2	1	4.3221	4.322080	0	3	1	2	2.2327	6.5548	6.5547	
0	3	0	3	2.9351	2.935110	0	3	1	2	3.6193	6.5544	6.5544	
0	2	2	1	2.3278	2.327828	0	3	1	2	4.2265	6.5543	6.5544	weak
0	4	2	2	6.8395	6.839377	0	3	3	1	0.6678	7.5073	7.5072	
0	4	0	4	8.3120	8.311898	0	3	3	1	-0.8046	7.5074	7.5073	
0	3	2	2	3.9710	3.970934	0	3	3	1	3.5361	7.5071	7.5070	
0	2	2	0	2.4213	2.421311	0	3	3	1	5.0860	7.5073	7.5073	
0	4	4	1	8.2954	8.295303	0	3	3	0	-0.7425	7.5529	7.5528	
0	3	2	1	4.3221	4.322080	0	3	3	0	3.2309	7.5530	7.5530	
0	2	2	1	2.3278	2.327828	0	3	3	0	5.2248	7.5526	7.5526	
0	5	0	5	6.8608	6.860890	0	4	1	4	0.4886	7.3494	7.3495	
0	4	2	3	6.0822	6.082186	0	4	1	4	1.2671	7.3493	7.3493	
0	3	0	3	2.9351	2.935110	0	4	1	4	4.4145	7.3496	7.3496	
0	5	2	4	8.6197	8.619788	0	4	1	3	-0.0580	8.5617	8.5618	component 1, blended
0	4	2	2	6.8395	6.839377	0	4	1	3	1.7222	8.5617	8.5616	
0	4	0	4	4.7032	4.703276	0	4	1	3	3.8584	8.5616	8.5617	
0	3	2	2	3.9710	3.970934	0	4	1	3	4.5908	8.5618	8.5618	
0	5	2	4	8.6197	8.619788	0	4	1	3	0.1185	8.7382	8.7383	component 2
0	4	2	2	6.8395	6.839377	0	4	1	3	1.8990	8.7385	8.7383	
0	4	0	4	4.7032	4.703276	0	4	1	3	4.0351	8.7383	8.7384	
0	3	2	2	3.9710	3.970934	0	4	1	3	4.7679	8.7389	8.7388	
0	5	4	1	11.3531	11.352928	0	4	3	2	-1.5563	9.7968	9.7966	
0	5	2	3	9.8561	9.856035	0	4	3	2	-0.0594	9.7967	9.7966	
0	4	4	1	8.2954	8.295303	0	4	3	2	1.5013	9.7967	9.7966	
0	4	2	3	6.0822	6.082186	0	4	3	2	3.7148	9.7970	9.7969	
0	3	2	1	4.3221	4.322080	0	4	3	2	5.4746	9.7967	9.7967	
0	5	4	2	11.2347	11.234530	0	4	3	1	-1.2089	10.0258	10.0256	
0	4	4	0	8.3120	8.311898	0	4	3	1	1.7136	10.0256	10.0255	
0	4	2	2	6.8395	6.839377	0	4	3	1	3.1866	10.0261	10.0260	
0	3	2	2	3.9710	3.970934	0	4	3	1	6.0545	10.0255	10.0255	
0	6	0	6	9.4096	9.409937	0	5	1	5	0.0883	9.4979	9.4983	
0	5	2	4	8.6197	8.619788	0	5	1	5	0.8783	9.4980	9.4981	
0	4	0	4	4.7032	4.703276	0	5	1	5	4.7953	9.4985	9.4985	
0	6	2	5	11.5602	11.560383	0	5	1	4	-0.3965	11.1637	11.1639	component 1
0	5	2	3	9.8561	9.856035	0	5	1	4	1.3077	11.1638	11.1637	
0	5	0	5	6.8608	6.860890	0	5	1	4	4.3030	11.1638	11.1639	
0	4	2	3	6.0822	6.082186	0	5	1	4	5.0816	11.1638	11.1638	
0	6	2	5	11.5602	11.560383	0	5	1	4	-0.2190	11.3412	11.3414	component 2
0	5	2	3	9.8561	9.856035	0	5	1	4	1.4853	11.3414	11.3413	
0	5	0	5	6.8608	6.860890	0	5	1	4	4.4808	11.3416	11.3416	
0	4	2	3	6.0822	6.082186	0	5	1	4	5.2591	11.3413	11.3413	
0	6	4	2	15.1375	15.137207	0	5	3	3	-2.5377	12.5998	12.5996	
0	6	2	4	13.2562	13.256243	0	5	3	3	-0.6579	12.5983	12.5984	blended
0	5	4	2	11.2347	11.234530	0	5	3	3	1.3649	12.5996	12.5994	
0	5	2	4	8.6197	8.619788	0	5	3	3	3.9795	12.5992	12.5993	
0	4	2	2	6.8395	6.839377	0	5	3	3	5.7600	12.5995	12.5994	
0	6	4	3	14.7434	14.743171	0	5	3	2	-1.5973	13.1461	13.1459	blended
0	5	4	1	11.3531	11.352928	0	5	3	2	1.7934	13.1465	13.1463	
0	5	2	3	9.8561	9.856035	0	5	3	2	3.2907	13.1468	13.1467	
0	4	4	1	8.2954	8.295303	0	5	3	2	4.8515	13.1469	13.1468	
0	4	2	3	6.0822	6.082186	0	5	3	2	7.0647	13.1469	13.1468	
0	6	6	0	17.7884	17.788197	0	5	5	1	-2.6539	15.1345	15.1343	
0	5	4	2	11.2347	11.234530	0	5	5	1	3.9001	15.1348	15.1346	
0	4	4	0	8.3120	8.311898	0	5	5	1	6.8224	15.1344	15.1343	
0	6	6	1	17.7865	17.786379	0	5	5	0	-2.6450	15.1415	15.1414	
0	5	4	1	11.3531	11.352928	0	5	5	0	3.7881	15.1412	15.1411	

Transitions are indicated with the lower state (") and upper state (') asymmetric-top quantum numbers, J , K_a and K_c . An offset of 2,950 cm⁻¹ has been subtracted from the line positions and upper state energies.

^aGround state energies calculated from parameters in ref. 23.

^bGround state energies from ref. 24.

Extended Data Table 3 | Continued from Extended Data Table 2

0 4 4 1	8.2954	8.295303	0 5 5 0	6.8458	15.1412	15.1411	
0 7 0 7	12.3503	12.350829	0 6 1 6	-0.3695	11.9808	11.9813	component 1
0 6 2 5	11.5602	11.560383	0 6 1 6	0.4209	11.9811	11.9812	
0 5 0 5	6.8608	6.860890	0 6 1 6	5.1207	11.9815	11.9816	
0 7 0 7	12.3503	12.350829	0 6 1 6	-0.2185	12.1318	12.1324	component 2, overlap
0 6 2 5	11.5602	11.560383	0 6 1 6	0.5715	12.1317	12.1319	
0 5 0 5	6.8608	6.860890	0 6 1 6	5.2709	12.1317	12.1318	
0 7 2 6	14.8950	14.895396	0 6 1 5	-0.7442	14.1508	14.1512	
0 6 2 4	13.2562	13.256243	0 6 1 5	0.8948	14.1510	14.1511	
0 6 0 6	9.4096	9.409937	0 6 1 5	4.7417	14.1513	14.1516	
0 5 2 4	8.6197	8.619788	0 6 1 5	5.5319	14.1516	14.1517	
0 7 2 5	17.0123	17.012438	0 6 3 4	-1.1530	15.8593	15.8594	
0 6 4 3	14.7434	14.743171	0 6 3 4	1.1160	15.8594	15.8591	
0 6 2 5	11.5602	11.560383	0 6 3 4	4.2986	15.8588	15.8590	
0 5 2 3	9.8561	9.856035	0 6 3 4	6.0030	15.8591	15.8591	
0 7 4 4	18.7676	18.767500	0 6 3 3				
0 6 4 2	15.1375	15.137207	0 6 3 3	1.7430	16.8805	16.8802	
0 6 2 4	13.2562	13.256243	0 6 3 3	3.6240	16.8802	16.8802	
0 5 4 2	11.2347	11.234530	0 6 3 3	5.6459	16.8806	16.8805	weak
0 5 2 4	8.6197	8.619788	0 6 3 3				
0 7 4 4	18.7676	18.767500	0 6 3 3				
0 6 4 2	15.1375	15.137207	0 6 3 3	1.6553	16.7928	16.7925	
0 6 2 4	13.2562	13.256243	0 6 3 3	3.5361	16.7923	16.7924	overlap
0 5 4 2	11.2347	11.234530	0 6 3 3				
0 5 2 4	8.6197	8.619788	0 6 3 3				
0 7 6 1	21.9487	21.948226	0 6 5 2	-3.2809	18.6678	18.6673	
0 6 6 1	17.7865	17.786379	0 6 5 2	0.8807	18.6672	18.6671	
0 6 4 3	14.7434	14.743171	0 6 5 2	3.9241	18.6675	18.6673	
0 5 4 1	11.3531	11.352928	0 6 5 2	7.3146	18.6677	18.6675	
0 7 6 2	21.9282	21.927772	0 6 5 1	-3.2034	18.7248	18.7244	
0 6 6 0	17.7884	17.788197	0 6 5 1	0.9366	18.7250	18.7248	
0 6 4 2	15.1375	15.137207	0 6 5 1	3.5872	18.7247	18.7244	
0 5 4 2	11.2347	11.234530	0 6 5 1	7.4901	18.7248	18.7247	
0 8 0 8	15.6827	15.683641	0 7 1 7	-0.7391	14.9436	14.9445	blended
0 7 2 6	14.8950	14.895396	0 7 1 7	0.0492	14.9442	14.9445	
0 6 0 6	9.4096	9.409937	0 7 1 7	5.5346	14.9442	14.9445	
0 8 2 7	18.6219	18.622625	0 7 1 6	-1.1604	17.4615	17.4622	?
0 7 2 5	17.0123	17.012438	0 7 1 6	0.4498	17.4621	17.4623	
0 7 0 7	12.3503	12.350829	0 7 1 6	5.1117	17.4620	17.4625	
0 6 2 5	11.5602	11.560383	0 7 1 6	5.9019	17.4621	17.4623	
0 8 2 6	21.1426	21.143107	0 7 3 5	-1.5966	19.5460	19.5465	blended
0 7 4 4	18.7676	18.767500	0 7 3 5	0.7782	19.5458	19.5457	
0 7 2 6	14.8950	14.895396	0 7 3 5	4.6502	19.5452	19.5456	
0 6 2 4	13.2562	13.256243	0 7 3 5	6.2896	19.5458	19.5458	
0 7 4 3	19.6150	19.614558	0 7 3 4	1.5684	21.1834	21.1830	tentative
0 7 2 5	17.0123	17.012438	0 7 3 4	4.1707	21.1830	21.1832	
0 6 4 3	14.7434	14.743171	0 7 3 4	6.4402	21.1836	21.1834	
0 7 4 3	19.6150	19.614558	0 7 3 4	1.3625	20.9775	20.9771	tentative
0 7 2 5	17.0123	17.012438	0 7 3 4	3.9658	20.9781	20.9782	
0 6 4 3	14.7434	14.743171	0 7 3 4	6.2355	20.9789	20.9787	
0 8 8 0	30.8313	30.831142	0 7 7 1	-4.5623	26.2690	26.2688	very weak, blended
0 6 6 0	17.7884	17.788197	0 7 7 1	8.4802	26.2686	26.2684	
0 8 8 1	30.8311	30.830978	0 7 7 0	-4.5623	26.2688	26.2687	very weak, blended
0 7 6 1	21.9487	21.948226	0 7 7 0	4.3211	26.2698	26.2693	
0 6 6 1	17.7865	17.786379	0 7 7 0	8.4830	26.2695	26.2694	
0 9 0 9	19.4070	19.408388	0 8 1 8	-1.1478	18.2592	18.2606	tentative
0 7 0 7	12.3503	12.350829	0 8 1 8	5.9091	18.2594	18.2599	" "

τ energy ordering only counts non-zero spin-weighted levels

m''	J''	τ''	m'	J'	τ'			
1 3 1	7.5474	7.545719	1 2 1		0.7904	8.3378	8.3361	
1 2 1	6.4605	6.459007	1 2 1					very weak
1 1 1	5.3918	5.390459	1 2 1		2.9457	8.3375	8.3361	
1 4 1	9.3853	9.383413	1 3 1		0.3800	9.7653	9.7634	
1 3 2	8.1993	8.197563	1 3 1					
1 2 1	6.4605	6.459007	1 3 1		3.3043	9.7648	9.7633	
1 5 1	11.5669	11.564673	1 4 1		0.0002	11.5671	11.5648	
1 4 2	10.5371	10.534994	1 4 1		1.0298	11.5669	11.5648	
1 3 2	8.1993	8.197563	1 4 1		3.3673	11.5666	11.5648	
1 3 1	7.5474	7.545719	1 4 1		4.0193	11.5667	11.5650	
1 6 1	14.1294	14.126832	1 5 1		-0.7197	13.4097	13.4071	
1 5 2	13.1369	13.134167	1 5 1		0.2729	13.4098	13.4071	
1 4 1	9.3853	9.383413	1 5 1		4.0240	13.4093	13.4074	
1 7 1	17.0794	17.076453	1 6 1		-1.0788	16.0006	15.9976	
1 5 1	11.5669	11.564673	1 6 1		4.4330	15.9999	15.9976	
1 8 1	20.4190	20.415573	1 7 1		-1.4281	18.9909	18.9874	
1 7 2	19.5036	19.500138	1 7 1		-0.5125	18.9911	18.9876	
1 6 1	14.1294	14.126832	1 7 1		4.8611	18.9905	18.9879	

The section below the black horizontal bar lists $|m| = 1$ transitions. Here, K_a and K_c are no longer used to label levels; instead, τ indicates different levels, in order of energy, with the same values of $|m|$ and J .

Extended Data Table 4 | Naphthalene ν_{29} band line list

J' Ka' Kc'	J'' Ka'' Kc''	Frequency	J' Ka' Kc'	J'' Ka'' Kc''	Frequency	J' Ka' Kc'	J'' Ka'' Kc''	Frequency
1 0 1	1 1 0	3064.5194	6 1 5	5 2 4	3064.9107	10 1 9	10 2 8	3064.2364
1 0 1	2 1 2	3064.4017	6 1 5	6 0 6	3064.8387	10 1 9	11 2 10	3063.8360
1 1 0	1 0 1	3064.6690	6 1 5	6 2 4	3064.4118	10 2 8	10 1 9	3064.9458
1 1 0	2 2 1	3064.2528	6 1 5	7 2 6	3064.0101	10 2 8	10 3 7	3064.3234
1 1 1	0 0 0	3064.7279	6 2 5	5 1 4	3065.1368	10 2 8	11 3 9	3063.6748
1 1 1	2 0 2	3064.5172	6 2 5	6 1 6	3064.9290	10 2 9	9 1 8	3065.2970
2 0 2	1 1 1	3064.6708	6 2 5	6 3 4	3064.2260	10 2 9	10 1 10	3065.1220
2 0 2	2 1 1	3064.5063	6 2 5	7 1 6	3064.1843	10 2 9	10 3 8	3064.1169
2 0 2	3 1 3	3064.3481	6 2 5	7 3 4	3063.7038	10 2 9	11 1 10	3063.8679
3 0 3	3 1 2	3064.4839	7 0 7	6 1 6	3065.0268	11 0 11	10 1 10	3065.2650
3 0 3	4 1 4	3064.2969	7 0 7	7 1 6	3064.2825	11 0 11	11 1 10	3064.0113
3 1 2	3 0 3	3064.7039	7 0 7	8 1 8	3064.0862	11 0 11	12 1 12	3063.8503
3 1 2	3 2 1	3064.4141	7 1 6	6 2 5	3065.0019	11 1 10	10 2 9	3065.3122
3 1 2	4 2 3	3064.1400	7 1 6	7 0 7	3064.9024	11 1 10	11 0 11	3065.1663
3 1 3	2 0 2	3064.8398	7 1 6	7 2 5	3064.3887	11 1 10	11 2 9	3064.1655
3 1 3	3 2 2	3064.3515	7 1 6	8 2 7	3063.9715	11 1 10	12 2 11	3063.7834
3 1 3	4 0 4	3064.3646	7 1 7	6 0 6	3065.0419	11 1 11	10 0 10	3065.2660
3 2 1	2 1 2	3065.0374	7 1 7	7 2 6	3064.2130	11 1 11	11 2 10	3063.9994
3 2 1	3 1 2	3064.7743	7 1 7	8 0 8	3064.0943	11 1 11	12 0 12	3063.8503
3 2 1	4 3 2	3063.9725	7 2 5	7 1 6	3064.7970	11 2 10	10 1 9	3065.3446
4 0 4	3 1 3	3064.8222	7 2 5	7 3 4	3064.3164	11 2 10	11 3 9	3064.0738
4 0 4	4 1 3	3064.4500	7 2 5	8 3 6	3063.7670	11 2 10	12 1 11	3063.8019
4 0 4	5 1 5	3064.2470	7 3 4	6 2 5	3065.4827	12 0 12	11 1 11	3065.3226
4 1 4	3 0 3	3064.8909	7 3 4	7 2 5	3064.8698	12 0 12	12 1 11	3063.9465
4 1 4	4 2 3	3064.3267	7 3 4	7 4 3	3064.1352	12 0 12	13 1 13	3063.7897
4 1 4	5 0 5	3064.2925	7 3 4	8 4 5	3063.5621	12 1 11	11 2 10	3065.3763
4 2 2	4 1 3	3064.7658	8 0 8	7 1 7	3065.0889	12 1 11	12 2 10	3064.0906
4 2 2	4 3 1	3064.2657	8 0 8	8 1 7	3064.2144	12 1 11	13 2 12	3063.7271
4 2 2	5 3 3	3063.9096	8 0 8	9 1 9	3064.0293	12 1 12	11 0 11	3065.3226
4 3 1	3 2 2	3065.2256	8 1 8	7 0 7	3065.0972	12 1 12	12 2 11	3063.9401
4 3 1	4 2 2	3064.9241	8 1 8	8 2 7	3064.1656	12 1 12	13 0 13	3063.7897
4 3 1	5 4 2	3063.7607	8 1 8	9 0 9	3064.0332	13 0 13	12 1 12	3065.3810
5 0 5	4 1 4	3064.8940	8 2 6	8 1 7	3064.8343	13 0 13	13 1 12	3063.8849
5 0 5	5 1 4	3064.4041	8 2 6	8 3 5	3064.3302	13 0 13	14 1 14	3063.7296
5 0 5	6 1 6	3064.1955	8 2 6	9 3 7	3063.7335	13 1 12	12 2 11	3065.4386
5 1 4	4 2 3	3064.8170	9 0 9	8 1 8	3065.1466	13 1 12	13 0 13	3065.2883
5 1 4	5 0 5	3064.7827	9 0 9	9 1 8	3064.1433	13 1 12	14 2 13	3063.6706
5 1 4	5 2 3	3064.4220	9 0 9	10 1 10	3063.9680	13 1 13	12 0 12	3065.3810
5 1 4	6 2 5	3064.0499	9 1 8	8 2 7	3065.1692	13 1 13	13 2 12	3063.8809
5 1 5	4 0 4	3064.9396	9 1 8	9 2 7	3064.2996	13 1 13	14 0 14	3063.7296
5 1 5	5 2 4	3064.2950	9 1 8	10 2 9	3063.8850	13 2 11	12 3 10	3065.4484
5 1 5	6 0 6	3064.2236	9 1 9	8 0 8	3065.1533	13 2 11	13 3 10	3064.2002
5 2 3	5 1 4	3064.7658	9 1 9	9 2 8	3064.1138	13 2 11	14 3 12	3063.5742
5 2 3	5 3 2	3064.2802	9 1 9	10 0 10	3063.9724	13 2 12	12 1 11	3065.4482
5 2 3	6 3 4	3063.8542	9 2 7	9 1 8	3064.8850	13 2 12	13 3 11	3063.9725
5 3 2	4 2 3	3065.3033	9 2 7	9 3 6	3064.3341	13 2 12	14 1 13	3063.6747
5 3 2	5 2 3	3064.9089	9 2 7	10 3 8	3063.7039	14 0 14	13 1 13	3065.4385
5 3 2	6 4 3	3063.6914	10 0 10	9 1 9	3065.2060	14 0 14	14 1 13	3063.8224
5 3 3	4 2 2	3065.2799	10 0 10	10 1 9	3064.0759	14 0 14	15 1 15	3063.6694
5 3 3	5 2 4	3064.9507	10 0 10	11 1 11	3063.9096	14 1 14	13 0 13	3065.4385
5 3 3	6 4 2	3063.6880	10 1 9	9 2 8	3065.2435	14 1 14	14 2 13	3063.8209
			10 1 9	10 0 10	3065.1025	14 1 14	15 0 15	3063.6694

155 assigned transitions, indicated by their upper and lower state (J, K_a, K_c) quantum numbers, are included in this list. Frequencies are given in cm^{-1} .

Ion-induced nucleation of pure biogenic particles

Jasper Kirkby^{1,2}, Jonathan Duplissy^{3,4}, Kamalika Sengupta⁵, Carla Frege⁶, Hamish Gordon², Christina Williamson^{1,†}, Martin Heinritzi^{1,7}, Mario Simon¹, Chao Yan³, João Almeida^{1,2}, Jasmin Tröstl⁶, Tuomo Nieminen^{3,4}, Ismael K. Ortega⁸, Robert Wagner³, Alexey Adamov³, Antonio Amorim⁹, Anne-Kathrin Bernhammer^{7,10}, Federico Bianchi^{6,11}, Martin Breitenlechner^{7,10}, Sophia Brilke¹, Xuemeng Chen³, Jill Craven¹², Antonio Dias², Sebastian Ehrhart^{1,2}, Richard C. Flagan¹², Alessandro Franchin³, Claudia Fuchs⁶, Roberto Guida², Jani Hakala³, Christopher R. Hoyle^{6,13}, Tuija Jokinen³, Heikki Junninen³, Juha Kangasluoma³, Jaeseok Kim^{14,†}, Manuel Krapf⁶, Andreas Kürten¹, Ari Laaksonen^{14,15}, Katrianne Lehtipalo^{3,6}, Vladimir Makhmutov¹⁶, Serge Mathot², Ugo Molteni⁶, Antti Onnela², Otsu Peräkylä³, Felix Piel¹, Tuukka Petäjä³, Arnaud P. Praplan³, Kirsty Pringle⁵, Alexandru Rap⁵, Nigel A. D. Richards^{5,17}, Ilona Riipinen¹⁸, Matti P. Rissanen³, Linda Rondo¹, Nina Sarnela³, Siegfried Schobesberger^{3,†}, Catherine E. Scott⁵, John H. Seinfeld¹², Mikko Sipilä^{3,4}, Gerhard Steiner^{3,7,19}, Yuri Stozhkov¹⁶, Frank Stratmann²⁰, Antonio Tomé²¹, Annele Virtanen¹⁴, Alexander L. Vogel², Andrea C. Wagner¹, Paul E. Wagner¹⁹, Ernest Weingartner⁶, Daniela Wimmer^{1,3}, Paul M. Winkler¹⁹, Penglin Ye²², Xuan Zhang¹², Armin Hansel^{7,10}, Josef Dommen⁶, Neil M. Donahue²², Douglas R. Worsnop^{3,14,23}, Urs Baltensperger⁶, Markku Kulmala^{3,4}, Kenneth S. Carslaw⁵ & Joachim Curtius¹

Atmospheric aerosols and their effect on clouds are thought to be important for anthropogenic radiative forcing of the climate, yet remain poorly understood¹. Globally, around half of cloud condensation nuclei originate from nucleation of atmospheric vapours². It is thought that sulfuric acid is essential to initiate most particle formation in the atmosphere^{3,4}, and that ions have a relatively minor role⁵. Some laboratory studies, however, have reported organic particle formation without the intentional addition of sulfuric acid, although contamination could not be excluded^{6,7}. Here we present evidence for the formation of aerosol particles from highly oxidized biogenic vapours in the absence of sulfuric acid in a large chamber under atmospheric conditions. The highly oxygenated molecules (HOMs) are produced by ozonolysis of α -pinene. We find that ions from Galactic cosmic rays increase the nucleation rate by one to two orders of magnitude compared with neutral nucleation. Our experimental findings are supported by quantum chemical calculations of the cluster binding energies of representative HOMs. Ion-induced nucleation of pure organic particles constitutes a potentially widespread source of aerosol particles in terrestrial environments with low sulfuric acid pollution.

It is thought that aerosol particles rarely form in the atmosphere without sulfuric acid^{3,4}, except in certain coastal regions where iodine oxides are involved⁸. Furthermore, ions are thought to be relatively unimportant in the continental boundary layer, accounting for only around 10% of particle formation⁵. Sulfuric acid derives from anthropogenic and volcanic sulfur dioxide emissions as well as dimethyl sulfide from marine biota. However, typical daytime sulfuric acid concentrations (10^5 – 10^7 cm⁻³, or 0.004–0.4 parts per trillion by volume (p.p.t.v.) at standard conditions) are too low for sulfuric acid and water alone to account for the particle formation rates observed in the lower atmosphere⁹, so additional vapours are required to stabilize any embryonic sulfuric acid clusters against evaporation. Base species such as amines can do this and can explain part of atmospheric particle

nucleation¹⁰. It is well established that oxidation products of volatile organic compounds (VOCs) are important for particle growth¹¹, but whether their role in the smallest particles is in nucleation or growth alone has remained ambiguous^{4,12,13}. Recently, however, it has been shown that oxidized organic compounds do indeed help to stabilize sulfuric acid clusters and probably play a major role in atmospheric particle nucleation^{6,14,15}. We refer to these compounds as HOMs (highly oxygenated molecules) rather than ELVOCs (extremely low-volatility organic compounds)¹⁶ because the measured compounds span a wide range of low volatilities.

Here we report atmospheric particle formation solely from biogenic vapours. The data were obtained at the CERN CLOUD chamber (Cosmics Leaving Outdoor Droplets; see Methods for experimental details) between October 2012 and November 2013. In contrast with other works that have reported organic particle formation without intentional addition of sulfuric acid^{6,7}, here we measure the cluster chemistry and the role of ions, and rule out contamination.

Precursor VOCs in the atmosphere arise predominantly from natural sources such as vegetation and largely comprise isoprene (C₅H₈), monoterpenes (C₁₀H₁₆), sesquiterpenes (C₁₅H₂₄) and diterpenes (C₂₀H₃₂). Here we have studied α -pinene (C₁₀H₁₆) because it is the most abundant monoterpene, often exceeding 50 p.p.t.v. in the continental boundary layer¹⁷. We oxidized α -pinene by exposure to ozone and also to hydroxyl radicals (OH·) produced from ozone photolysis and secondary reactions. To measure the relative importance of these oxidants we also performed a few pure ozonolysis experiments (in which we removed OH· with a 0.1% H₂ scavenger) and a few pure hydroxyl experiments (in which we generated OH· by photolysis of gas-phase nitrous acid, HONO). Two nitrate chemical ionization atmospheric pressure interface time-of-flight (CI-API-TOF) mass spectrometers measured neutral gas-phase compounds in the chamber (H₂SO₄ and HOMs). Therefore, for this study, HOMs are implicitly defined as oxidized organic compounds that can be detected by a nitrate CI-API-TOF; related molecules with a lower oxidation state

¹Goethe University Frankfurt, Institute for Atmospheric and Environmental Sciences, 60438 Frankfurt am Main, Germany. ²CERN, CH-1211 Geneva, Switzerland. ³Department of Physics, University of Helsinki, FI-00014 Helsinki, Finland. ⁴Helsinki Institute of Physics, University of Helsinki, FI-00014 Helsinki, Finland. ⁵School of Earth and Environment, University of Leeds, Leeds LS2 9JT, UK. ⁶Paul Scherrer Institute, Laboratory of Atmospheric Chemistry, CH-5232 Villigen, Switzerland. ⁷Institute for Ion and Applied Physics, University of Innsbruck, 6020 Innsbruck, Austria. ⁸Onera—The French Aerospace Lab, F-91123 Palaiseau, France. ⁹SIM, University of Lisbon, 1849-016 Lisbon, Portugal. ¹⁰Ionicon Analytik GmbH, 6020 Innsbruck, Austria. ¹¹Institute for Atmospheric and Climate Science, ETH Zurich, CH-8092 Zurich, Switzerland. ¹²Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125, USA. ¹³WSL Institute for Snow and Avalanche Research SLF, CH-7260 Davos, Switzerland. ¹⁴University of Eastern Finland, FI-70211 Kuopio, Finland. ¹⁵Finnish Meteorological Institute, FI-00010 Helsinki, Finland. ¹⁶Solar and Cosmic Ray Research Laboratory, Lebedev Physical Institute, 119991 Moscow, Russia. ¹⁷University of Leeds, National Centre for Earth Observation, Leeds LS2 9JT, UK. ¹⁸Department of Applied Environmental Science, University of Stockholm, SE-10961 Stockholm, Sweden. ¹⁹Faculty of Physics, University of Vienna, 1090 Vienna, Austria. ²⁰Leibniz Institute for Tropospheric Research, 04318 Leipzig, Germany. ²¹University of Beira Interior, 6201-001 Covilhã, Portugal. ²²Center for Atmospheric Particle Studies, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA. ²³Aerodyne Research Inc., Billerica, Massachusetts 01821, USA. †Present addresses: CIRES, University of Colorado Boulder, Boulder, Colorado 80309, USA (C.W.); Arctic Research Center, Korea Polar Research Institute, Incheon 406-840, South Korea (J. Kim); Department of Atmospheric Sciences, University of Washington, Seattle, Washington 98195, USA (S.S.).

or different functional groups could be present in the chamber, but undetected by our nitrate chemical ionization set-up.

Before starting measurements, we carefully cleaned the CLOUD chamber (see Methods) and established extremely low contaminant concentrations: at 38% relative humidity and 278 K, the contaminants were below the detection limit for SO_2 (<15 p.p.t.v.) and H_2SO_4 ($<5 \times 10^4 \text{ cm}^{-3}$), and total organics (largely comprising high volatility C_1 – C_3 compounds) were below 150 p.p.t.v. Contaminants with a high proton affinity or a high gas-phase acidity can be detected as ions by the API-TOF operating in positive or negative mode, respectively, even at neutral molecule concentrations as low as 10^4 cm^{-3} . The API-TOF measured contaminant $\text{C}_5\text{H}_5\text{NH}^+$ (protonated pyridine) and contaminant NO_3^- to be the dominant positive and negative ions, respectively, before we added any trace gases to the chamber other than water vapour and ozone (Extended Data Fig. 1a, b). Despite its higher gas-phase acidity, we detected contaminant HSO_4^- at only 1% of the NO_3^- signal (Extended Data Fig. 1b), ruling out any contribution of sulfuric acid to the nucleation measurements. From previous studies and molecular analysis of the charged clusters (see below), the most abundant positive ion is likely to be contaminant ammonium (NH_4^+), but its mass is below the acceptance cut-off of the API-TOF as operated in this study.

Within a few minutes of the initial exposure of α -pinene to O_3 in the chamber, we detected gas-phase HOM monomers and dimers (Fig. 1a). Particles appeared shortly afterwards (Fig. 1b). HOM monomers (denoted E_1) broadly comprise highly oxidized $\text{C}_{8-10}\text{H}_{14,16}\text{O}_{6-12}$ species with an oxygen-to-carbon ratio (O/C) above about 0.6. HOM dimers (E_2) are two covalently bound monomers (see below), which generally have lower oxygen-to-carbon ratios, but, almost certainly, a lower volatility. For the present study we define E_1 (E_2) to be the summed HOM peaks in the mass/charge range $m/z = 235$ – 424 Th (425 – 625 Th), where $1 \text{ Th} = 1 \text{ Da}/e$ and e is the elementary charge. This definition excludes peaks in the E_1 mass band distinguished by an odd H number ($\text{C}_{10}\text{H}_{15}\text{O}_{6,8,10,12}$), which we assign to the RO_2^\cdot peroxy radical. These m/z values include a contribution of 62 Th due to the NO_3^- ion from the CI-API-TOF ionizer. We define the total HOMs as the sum $\text{RO}_2^\cdot + \text{E}_1 + \text{E}_2$.

We measure high HOM molar yields (Extended Data Fig. 2): approximately 1.2% per hydroxyl radical (OH^\cdot) reaction with α -pinene, 3.2% per ozone reaction with α -pinene, and 2.9% from pure ozonolysis. We find a high E_2 yield from ozonolysis (10%–20% of total HOMs), but negligible E_2 yield from hydroxyl-initiated oxidation. Neutral trimers are close to the detection limit of the CI-API-TOF (below 0.1% of total HOMs). High yields of these same HOMs have previously been reported^{6,16}, although our ozonolysis yields are less than half those of ref. 16. For our experiments, α -pinene was in the range 0.1–2 parts per billion by volume (p.p.b.v.), with 20–40 p.p.b.v. of O_3 . The OH^\cdot concentrations were $(0.5\text{--}0.8) \times 10^6 \text{ cm}^{-3}$ during ozonolysis experiments, and $(0.4\text{--}2) \times 10^5 \text{ cm}^{-3}$ during pure hydroxyl experiments with 0.5–3 p.p.b.v. of HONO.

This remarkably fast production of HOMs is likely to proceed via an autooxidation mechanism involving peroxy radicals^{16,18–20} (Extended Data Fig. 3). There is simply insufficient time for oxidation to proceed in multiple steps through stable intermediate molecules. Here, initial ozonolysis of an α -pinene molecule proceeds via a Criegee intermediate and further steps to form an RO_2^\cdot radical, followed by several repeated cycles of intramolecular H abstraction and O_2 addition to re-form a new RO_2^\cdot radical. We measure an RO_2^\cdot fraction of total HOMs between 15% and 1% for HOMs from 0.1 p.p.t.v. to 10 p.p.t.v., respectively. A combination reaction of differently oxidized peroxy radicals explains the rapid high yield of covalently bound E_2 . The negligible E_2 yield from hydroxyl-initiated oxidation could result from additional NO_x chemistry that terminates the peroxy radicals before they can combine. Our theoretical calculations further indicate that E_2 must be covalently bound because the neutral molecular cluster formed from two monomers (denoted E_1E_1) is expected to be unstable (see below).

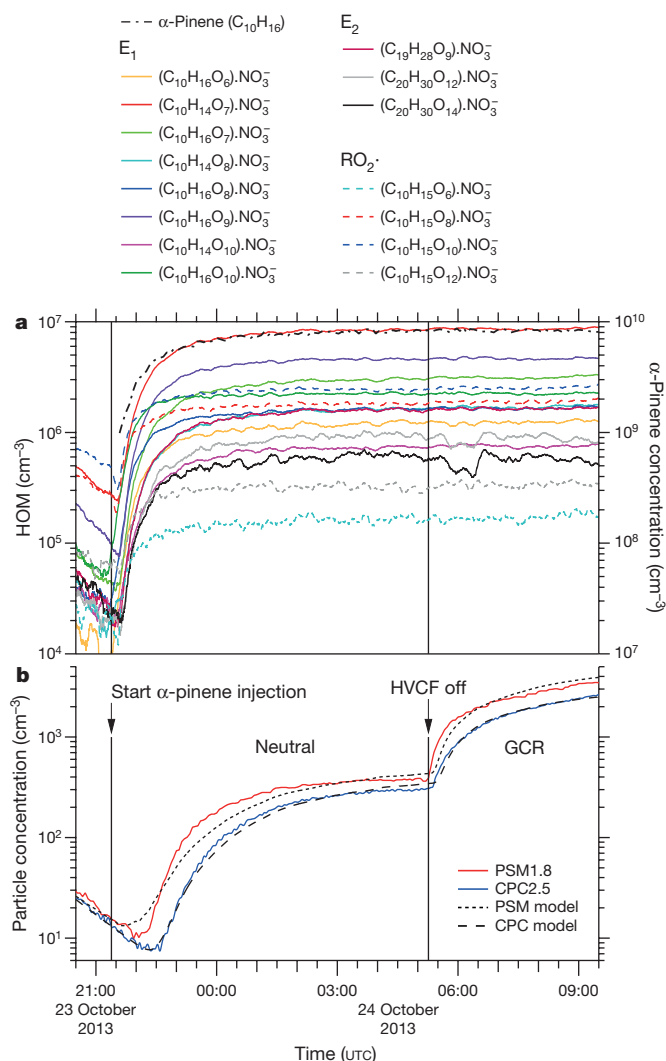


Figure 1 | Evolution of HOMs and particles during a typical run.

a, Evolution of selected HOM monomers (E_1), dimers (E_2) and peroxy radicals (RO_2^\cdot) at 300 p.p.t.v. α -pinene, 33 p.p.b.v. O_3 , zero H_2 or HONO, 38% relative humidity, 278 K and $[\text{H}_2\text{SO}_4] < 5 \times 10^4 \text{ cm}^{-3}$ (the same run as shown in Extended Data Fig. 4). The HOMs start to appear soon after the first injection of α -pinene into the chamber at 21:22, 23 October 2013. A HOM monomer is a highly oxygenated molecule derived from α -pinene ($\text{C}_{10}\text{H}_{16}$), and a HOM dimer is a covalently bound pair of monomers. Peroxy radicals are identified by an odd H number. The HOMs are charged with an NO_3^- ion in the CI-API-TOF mass spectrometer. The systematic scale uncertainty on the HOM concentrations is $+80\%/ -45\%$. **b**, Evolution of the particle number concentrations measured in the PSM1.8 (red curve) and CPC2.5 (blue curve) particle counters. The high-voltage clearing field (HVCF) was switched off at 05:16, 24 October 2013, marking the transition from neutral (ion-free) to GCR conditions in the chamber. A sharp increase in the rate of particle formation is seen, due to ion-induced nucleation of pure biogenic particles. However, no change occurs in the HOM concentrations (**a**), because these are predominantly neutral gas-phase molecules. The dotted and dashed curves in **b** show the PSM1.8 and CPC2.5 distributions, respectively, simulated for this run with the AEROCLOUD kinetic model, which is used to derive the experimental nucleation rates (see Methods).

We measured nucleation rates under neutral (J_n), Galactic cosmic ray (GCR; J_{gcr}) and π^+ beam (J_π) conditions, corresponding to ion-pair concentrations of around 0 cm^{-3} , 700 cm^{-3} and $3,000 \text{ cm}^{-3}$, respectively. This range spans atmospheric ion concentrations between ground level and 15-km altitude. The nucleation rate J_n describes the neutral rate alone, whereas J_{gcr} and J_π describe the sum of the neutral and ion-induced rates, $J_n + J_{\text{ion}}$. We determine the nucleation rates at

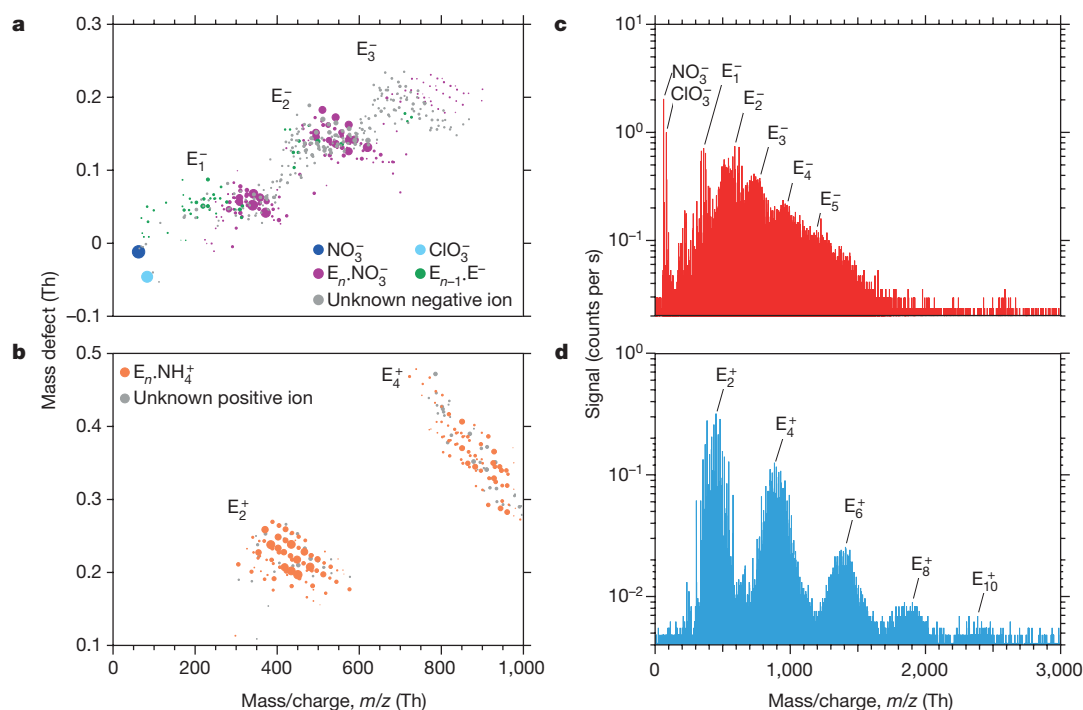


Figure 2 | Molecular composition and mass spectra of charged clusters during GCR nucleation events without sulfuric acid. **a, b,** Cluster mass defect (difference from integer mass) versus m/z of negatively (**a**) and positively (**b**) charged clusters measured with the API-TOF at 240 p.p.t.v. α -pinene, 34 p.p.b.v. O_3 , zero H_2 or HONO, 38% relative humidity, 278 K and $[H_2SO_4]$ below the detection limit ($5 \times 10^4 \text{ cm}^{-3}$). The values of J_{GCR} and total HOMs concentration are, respectively, $3.4 \text{ cm}^{-3} \text{ s}^{-1}$ and $1.7 \times 10^7 \text{ cm}^{-3}$ (**a**), and $3.3 \text{ cm}^{-3} \text{ s}^{-1}$ and $2.4 \times 10^7 \text{ cm}^{-3}$ (**b**). The mass bands are labelled according to the number of HOM monomer units in the cluster, E_n . Each circle represents a distinct molecular composition and its area represents the counts per second. The most highly oxidized compounds are located at the lower right-hand edge of each band.

1.7-nm mobility diameter, at which size a particle is generally considered to be stable against evaporation. To determine the nucleation rates, we fit the time-dependent particle concentrations with a numerical model that treats particle nucleation and growth kinetically at the molecular level (an example is shown in Fig. 1b; see Methods for further details).

A typical run sequence (Extended Data Fig. 4) begins by establishing ion-free conditions with a high-voltage clearing field and introducing α -pinene to the chamber, where it mixes with ozone. Particles then start to form and, after measuring J_n at steady-state α -pinene concentration, we turn off the high voltage and measure J_{GCR} under otherwise identical chamber conditions. A sharp enhancement of particle formation is seen when the high voltage was turned off (Extended Data Fig. 4b, e), due to ion-induced nucleation of both charge signs (Extended Data Figs 4c, d and 5).

Figure 2 shows the molecular composition and mass spectra of negatively and positively charged ions, monomers, dimers and clusters during ion-induced nucleation events. The dominant core ions in the clusters are identified as NH_4^+ , NO_3^- and E^- . Here E^- is inferred for negatively charged ions or clusters that contain only C, H and O; the E^- ion corresponds to a HOM of high gas-phase acidity. In contrast to negative clusters, the positive clusters nucleate only with dimers, producing distinct mass bands that are detected up to E_{10} in the API-TOF (Fig. 2c, d). This indicates the importance of dimers for pure biogenic nucleation. Dimers are expected to be less volatile than monomers, owing primarily to higher molecular weight, but also to additional functional groups. Our previously described definition for neutral gas-phase HOMs encompasses compounds with a wide range

The dark blue circle represents NO_3^- ions; the light blue circle represents ClO_3^- ions. Clusters with fully identified molecular composition are coloured according to their core ion: purple (NO_3^-), green (E^-) or orange (NH_4^+). Grey circles are unidentified clusters. **c, d,** Mass spectra from the same events for negative (**c**) and positive (**d**) clusters up to $m/z = 3,000$ Th. A particle of 1.7-nm mobility diameter has a mass of about 1,200 Th. The 'Nessie' plot (**d**) shows that positive-ion-induced nucleation involves HOM dimers alone ($E_1NH_4^+$ clusters are not seen owing to instrument tuning). The decreasing signal amplitude at larger masses is due to the lower concentration and decreasing detection efficiency of the API-TOF mass spectrometer (the efficiency versus m/z depends on the instrument tune and polarity).

of low volatilities^{19,21}, of which only a subset drive nucleation (ELVOCs, which comprise about 36% of measured total HOMs²¹). From the strong ion enhancement of nucleation we conclude that the API-TOF mass peaks above the dimer in Fig. 2 are clusters of ELVOC monomers and dimers. Although we can precisely determine their molecular composition ($C_xH_yO_z$), we can only infer their specific structure and functional groups.

We show the experimental neutral and GCR nucleation rates in Fig. 3 over the total HOMs range 0.1–10 p.p.t.v., which spans the range of atmospheric interest. Below 1 p.p.t.v. HOM, ionization at ground-level GCR intensities enhances the nucleation rate by between one and two orders of magnitude compared with neutral nucleation. At higher concentrations, the neutral and GCR nucleation rates converge because the ion-induced rate, J_{ion} , reaches the limit set by the GCR total ion production rate ($3.4 \text{ cm}^{-3} \text{ s}^{-1}$). Positive and negative clusters nucleate at comparable rates (an example is shown in Extended Data Fig. 5). Relative humidity has little effect on J_{GCR} over the range 6%–80% relative humidity, whereas J_n increases substantially at higher relative humidity (Extended Data Fig. 6).

The large GCR enhancement indicates that biogenic molecular clusters are relatively unstable unless an ion is present. A charged cluster is also likely to experience higher collision rates with HOMs because they are expected to have high electric polarizability and, depending on their structure, large dipole moments. We further investigated the dependence on ion species by adding small amounts of SO_2 to the chamber, up to around 1,000 p.p.t.v. When $[H_2SO_4]$ exceeds about $1 \times 10^5 \text{ cm}^{-3}$, the major negative ion species shift to HSO_4^- , SO_5^- and SO_4^- (Extended Data Fig. 1c), owing to their lower proton affinity (higher gas-phase

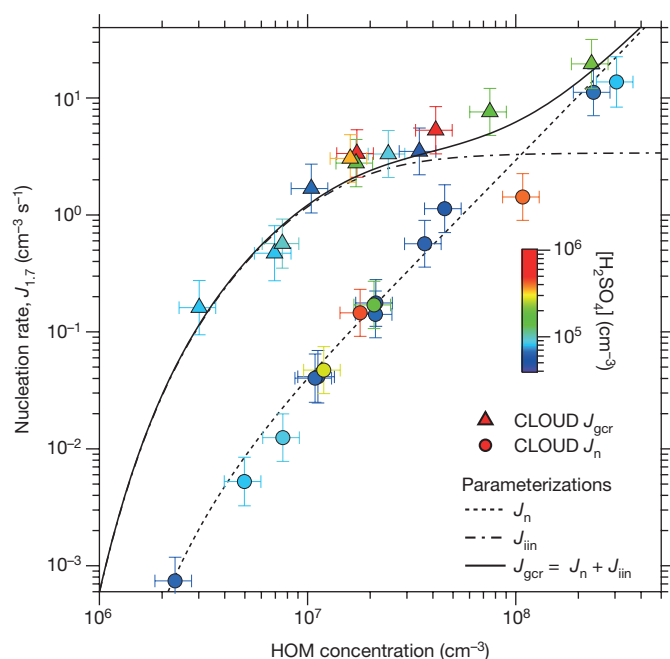


Figure 3 | Pure biogenic nucleation rates versus HOM concentration. Neutral (J_n ; circles) and GCR (J_{gr} ; triangles) nucleation rates versus total HOMs concentration ($\text{RO}_2 + \text{E}_1 + \text{E}_2$). The fraction of total HOMs that participate in nucleation (ELVOCs) is about 36% (ref. 21). The experimental conditions are 10–1,300 p.p.t.v. α -pinene (for measurements below $J_{1.7} = 10 \text{ cm}^{-3} \text{ s}^{-1}$), 30–35 p.p.b.v. O_3 , zero H_2 or HONO, 38% relative humidity, 278 K and $< 8 \times 10^5 \text{ cm}^{-3} \text{ H}_2\text{SO}_4$. The colour scale shows $[\text{H}_2\text{SO}_4]$; purple and blue points correspond to contaminant level (below the detection threshold); other colours correspond to measurements after SO_2 was added to the chamber. The fitted curves show parameterizations (described in Methods) for J_n (dashed), J_{gr} (solid) and ion-induced nucleation ($J_{iin} = J_{gr} - J_n$; dot-dashed). The J_{iin} parameterization assumes that the nucleation rate falls steeply at HOM concentrations below the experimental measurements, following a similar slope to that for J_n . The bars indicate 1σ total errors, although the overall systematic scale uncertainty of $+80\%/-45\%$ on the HOM concentration is not shown.

acidity) than contaminant compounds. However, the nucleation rates with sulfur ion species remain unchanged (Fig. 3). Taken together, our observations therefore show that ubiquitous ion species can stabilize embryonic biogenic clusters. However, we do not observe chlorine in nucleating clusters, even though contaminant chlorine ion species are present (Fig. 2 and Extended Data Fig. 1), which indicates that not all ions have a suitable chemical structure to bond strongly with the oxidized organic compounds²².

Figure 4 shows the CLOUD biogenic nucleation rates extended to $[\text{H}_2\text{SO}_4] = 6 \times 10^6 \text{ cm}^{-3}$ and compared with atmospheric boundary-layer observations^{3,4,23,24}. Biogenic nucleation rates show no significant dependence on sulfuric acid concentration over this range (that is, within the experimental measurement errors, the nucleation rate is consistent with zero dependency on sulfuric acid concentration). This finding sharply contrasts with base-stabilized nucleation of sulfuric acid in the presence of ammonia⁹ or amines¹⁰, where nucleation rates at 1.7 nm show a steep dependency on $[\text{H}_2\text{SO}_4]$ above 10^6 cm^{-3} . Comparison of the atmospheric observations (Fig. 4) with our measurements therefore suggests that nucleation in the lower atmosphere may involve a mixture of two distinct mechanisms. The first, which is more important in polluted environments, involves nucleation of sulfuric acid and water together with a combination of amines or ammonia with oxidized organics, and has a strong dependence on sulfuric acid. The second, which is more important in pristine environments, involves nucleation of pure organic particles and depends on only oxidized organics and ions.

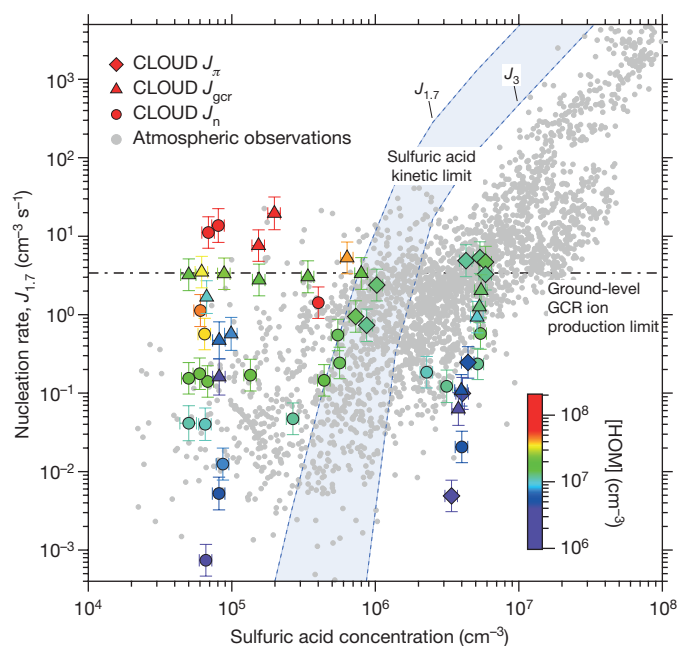


Figure 4 | Experimental and atmospheric nucleation rates versus H_2SO_4 concentration. CLOUD measurements of the neutral (J_n ; circles), GCR (J_{gr} ; triangles) and π beam (J_{pi} ; diamonds) biogenic nucleation rates at 1.7 nm ($J_{1.7}$) versus $[\text{H}_2\text{SO}_4]$. The CLOUD experimental conditions are 10–1,300 p.p.t.v. α -pinene (for measurements below $J_{1.7} = 10 \text{ cm}^{-3} \text{ s}^{-1}$), 25–35 p.p.b.v. O_3 , zero H_2 or HONO, 20%–40% relative humidity and 278 K. Measurements below $1 \times 10^5 \text{ cm}^{-3}$ for $[\text{H}_2\text{SO}_4]$ are near to the detection limit of the CI-API-TOF and should be considered as upper-estimates (to avoid overlap, some data points at the H_2SO_4 detection limit are displaced by up to $1 \times 10^4 \text{ cm}^{-3}$). The total HOMs concentration from α -pinene oxidation is indicated by the colour scale. Observations of particle formation in the atmospheric boundary layer (mainly at 3-nm threshold size) are indicated by small grey circles^{3,4,23,24}. Following convention, the H_2SO_4 concentration refers to monomers alone; that is, H_2SO_4 bound in molecular clusters is not included. The kinetic upper limit on sulfuric acid nucleation is indicated by the blue band, which is bounded by dashed lines indicating $J_{1.7}$ and J_3 . This band assumes the CLOUD condensation sink, which is comparable to that of a pristine atmosphere. The upper limit on J_{iin} from the GCR ion-pair production rate at ground level is indicated by the dot-dashed line. The bars indicate 1σ total errors, although the overall $+50\%/-33\%$ systematic scale uncertainty on $[\text{H}_2\text{SO}_4]$ is not shown.

To gain further insight into the stability of initial neutral and charged clusters of highly oxidized biogenic molecules, we calculated their Gibbs free energies of formation, ΔG , using quantum chemical methods (see Methods). For this study we chose $\text{C}_{10}\text{H}_{14}\text{O}_7$ and $\text{C}_{20}\text{H}_{30}\text{O}_{14}$ as E_1 and E_2 surrogates, respectively (Extended Data Fig. 7). We observe these compounds both in the gas (Fig. 1) and particle phases in the CLOUD chamber. We show proposed formation mechanisms and structures^{19,20} in Extended Data Fig. 3. Our calculations, summarized in Extended Data Table 1 and Extended Data Fig. 8, confirm that ELVOC clusters formed with an E_1^- , HSO_4^- , NO_3^- or NH_4^+ ion are expected to be stable (that is, their growth rate exceeds the evaporation rate) at around 0.1 p.p.t.v. ELVOC, or below. In contrast, the initial neutral clusters are weakly bound and so neutral nucleation is expected to be weaker. Although limited to a single surrogate pair, our theoretical calculations thus provide independent support for the experimental measurements.

Comparisons with atmospheric observations should be considered as preliminary because our measurements were made at only one temperature, with a single monoterpene, in the absence of isoprene and mostly in the absence of NO_x , which can influence HOM yields. Nevertheless, our results may provide fresh insights into several seemingly disparate

phenomena associated with low atmospheric concentrations of sulfuric acid. First, pure HOM nucleation could provide a mechanism to account for nucleation-mode particles observed at night-time, under low- $[\text{H}_2\text{SO}_4]$ conditions^{25,26}. Second, although observations are rare, nucleation-mode particles are seen in the Amazon²⁷, where SO_2 levels are extremely low (20–30 p.p.t.v.). Peak particle concentrations often occur at sunrise and sunset²⁷, and appear to be associated with rain, which reduces the aerosol condensation sink and may generate high ion concentrations by evaporation of charged droplets at the Rayleigh limit. Third, pure biogenic nucleation could explain new particle formation observed in the upper troposphere in cloud outflows depleted of SO_2 , such as over the Amazon^{27–29}. Low-solubility biogenic precursor vapours can be efficiently convected inside clouds to high altitudes where HOMs will form in the cloud outflows on exposure to oxidants, and nucleation is likely to be enhanced by the low temperatures. Fourth, since high HOM yields are also found from other organic compounds with an endocyclic double bond such as cyclohexene¹⁶, pure HOM nucleation involving anthropogenic organic precursors could be expected when $[\text{H}_2\text{SO}_4]$ is low³⁰. Finally, ion-induced pure biogenic nucleation might shed new light on the long-standing question of a physical mechanism for solar-climate variability in the pristine pre-industrial climate^{31,32}.

Direct observational evidence of pure biogenic nucleation has not been reported so far, owing to atmospheric pollution or lack of suitable instrumentation. The pure biogenic mechanism is likely to dominate nucleation in pristine terrestrial regions such as tropical rainforests or at higher altitudes above forests in convective cloud outflows. Pure biogenic nucleation might also take place over forested areas at high northern latitudes during periods of especially low pollution. Identification of pure biogenic nucleation in the atmosphere will require simultaneous measurements with several newly developed mass spectrometers, API-TOF (for molecular composition of ions and nucleating charged clusters) and CI-API-TOF (gas-phase HOMs and H_2SO_4), together with standard instruments such as low-threshold particle counters, PTR-TOF (precursor organic vapours) and NAIS (size spectra of ions and charged particles).

In summary, we find that highly oxidized organic compounds play a role in atmospheric particle nucleation comparable to that of sulfuric acid; together with a suitable stabilizing agent, each has sufficiently low volatility to form new particles in the lower atmosphere at vapour concentrations near 10^7 cm^{-3} . The stabilizing agent for pure biogenic particles is a suitable ion, whereas for sulfuric acid particles the stabilizing agents are amines, or ammonia with oxidized organics. Ion-induced nucleation of pure biogenic particles may have important consequences for pristine climates because it provides a mechanism by which nature produces particles without pollution. This could raise the baseline aerosol state of the pristine pre-industrial atmosphere and so could reduce the estimated anthropogenic radiative forcing from increased aerosol-cloud albedo over the industrial period.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 July 2015; accepted 16 March 2016.

1. Boucher, O. *et al.* in *Climate Change 2013: The Physical Science Basis. Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Stocker, T. F. *et al.*) 571–658 (Cambridge Univ. Press, 2013).
2. Merikanto, J., Spracklen, D. V., Mann, G. W., Pickering, S. J. & Carslaw, K. S. Impact of nucleation on global CCN. *Atmos. Chem. Phys.* **9**, 8601–8616 (2009).
3. Kuang, C., McMurry, P. H., McCormick, A. V. & Eisele, F. L. Dependence of nucleation rates on sulfuric acid vapor concentration in diverse atmospheric locations. *J. Geophys. Res. Atmos.* **113**, D10209 (2008).
4. Kulmala, M. *et al.* Direct observations of atmospheric aerosol nucleation. *Science* **339**, 943–946 (2013).
5. Hirsikko, A. *et al.* Atmospheric ions and nucleation: a review of observations. *Atmos. Chem. Phys.* **11**, 767–798 (2011).

6. Zhao, J., Ortega, J., Chen, M., McMurry, P. H. & Smith, J. N. Dependence of particle nucleation and growth on high-molecular-weight gas-phase products during ozonolysis of α -pinene. *Atmos. Chem. Phys.* **13**, 7631–7644 (2013).
7. Gao, S. *et al.* Low-molecular-weight and oligomeric components in secondary organic aerosol from the ozonolysis of cycloalkenes and α -pinene. *J. Phys. Chem. A* **108**, 10147–10164 (2004).
8. O'Dowd, C. D. *et al.* Marine aerosol formation from biogenic iodine emissions. *Nature* **417**, 632–636 (2002).
9. Kirkby, J. *et al.* Role of sulphuric acid, ammonia and galactic cosmic rays in atmospheric aerosol nucleation. *Nature* **476**, 429–433 (2011).
10. Almeida, J. *et al.* Molecular understanding of sulphuric acid–amine particle nucleation in the atmosphere. *Nature* **502**, 359–363 (2013).
11. Riipinen, I. *et al.* Organic condensation: a vital link connecting aerosol formation to cloud condensation nuclei (CCN) concentrations. *Atmos. Chem. Phys.* **11**, 3865–3878 (2011).
12. Zhang, R. *et al.* Atmospheric new particle formation enhanced by organic acids. *Science* **304**, 1487–1490 (2004).
13. Metzger, A. *et al.* Evidence for the role of organics in aerosol particle formation under atmospheric conditions. *Proc. Natl Acad. Sci. USA* **107**, 6646–6651 (2010).
14. Schobesberger, S. *et al.* Molecular understanding of atmospheric particle formation from sulfuric acid and large oxidized organic molecules. *Proc. Natl Acad. Sci. USA* **110**, 17223–17228 (2013).
15. Riccobono, F. *et al.* Oxidation products of biogenic emissions contribute to nucleation of atmospheric particles. *Science* **344**, 717–721 (2014).
16. Ehn, M. *et al.* A large source of low-volatility secondary organic aerosol. *Nature* **506**, 476–479 (2014).
17. Guenther, A. B. *et al.* The Model of Emissions of Gases and Aerosols from Nature version 2.1 (MEGAN2.1): an extended and updated framework for modeling biogenic emissions. *Geosci. Model Dev.* **5**, 1471–1492 (2012).
18. Crounse, J. D., Nielsen, L. B., Jørgensen, S., Kjaergaard, H. G. & Wennberg, P. O. Autoxidation of organic compounds in the atmosphere. *J. Phys. Chem. Lett.* **4**, 3513–3520 (2013).
19. Zhang, X. *et al.* Formation and evolution of molecular products in α -pinene secondary organic aerosol. *Proc. Natl Acad. Sci. USA* **112**, 14168–14173 (2015).
20. Kurtén, T. *et al.* Computational study of hydrogen shifts and ring-opening mechanisms in α -pinene ozonolysis products. *J. Phys. Chem. A* **119**, 11366–11375 (2015).
21. Tröstl, J. *et al.* The role of low-volatility organic compounds in initial particle growth in the atmosphere. *Nature* **533**, <http://dx.doi.org/10.1038/nature18271> (2016).
22. Kathmann, S. M., Schenter, G. K. & Garrett, B. C. Ion-induced nucleation: the importance of chemistry. *Phys. Rev. Lett.* **94**, 116104 (2005).
23. Paasonen, P. *et al.* On the roles of sulphuric acid and low-volatility organic vapours in the initial steps of atmospheric new particle formation. *Atmos. Chem. Phys.* **10**, 11223–11242 (2010).
24. Sihto, S.-L. *et al.* Atmospheric sulphuric acid and aerosol formation: implications from atmospheric measurements for nucleation and early growth mechanisms. *Atmos. Chem. Phys.* **6**, 4079–4091 (2006).
25. Suni, T. *et al.* Formation and characteristics of ions and charged aerosol particles in a native Australian eucalyptus forest. *Atmos. Chem. Phys.* **8**, 129–139 (2008).
26. Lee, S.-H. *et al.* Observations of nighttime new particle formation in the troposphere. *J. Geophys. Res. Atmos.* **113**, D10210 (2008).
27. Martin, S. T. *et al.* Sources and properties of Amazonian aerosol particles. *Rev. Geophys.* **48**, RG2002 (2010).
28. Kulmala, M. *et al.* Deep convective clouds as aerosol production engines: role of insoluble organics. *J. Geophys. Res. Atmos.* **111**, D17202 (2006).
29. Ekman, A. M. L. *et al.* Do organics contribute to small particle formation in the Amazonian upper troposphere? *Geophys. Res. Lett.* **35**, L17810 (2008).
30. Bianchi, F. *et al.* New particle formation in the free troposphere: a question of chemistry and timing. *Science* **352**, <http://dx.doi.org/10.1126/science.aad5456> (2016).
31. Herschel, W. Observations tending to investigate the nature of the Sun, in order to find the causes or symptoms of its variable emission of light and heat; with remarks on the use that may possibly be drawn from solar observations. *Philos. Trans. R. Soc. Lond.* **91**, 265–318 (1801).
32. Kirkby, J. Cosmic rays and climate. *Surv. Geophys.* **28**, 333–375 (2007).

Acknowledgements We thank CERN for supporting CLOUD with important technical and financial resources, and for providing a particle beam from the CERN Proton Synchrotron. We also thank P. Carrie, L.-P. De Menezes, J. Dumollard, F. Josa, I. Krasin, R. Kristic, A. Laassiri, O. S. Maksumov, B. Marichy, H. Martinati, S. V. Mizin, R. Sitals, A. Wasem and M. Wilhelmsson for their contributions to the experiment. We thank the CSC Centre for Scientific Computing in Espoo, Finland for computer time. This research has received funding from the EC Seventh Framework Programme (Marie Curie Initial Training Network MC-ITN CLOUD-TRAIN no. 316662, EU Horizon 2020 Marie Curie grant no. 656994, ERC-Consolidator grant NANODYNAMITE no. 616075 and ERC-Advanced grant ATMNUCLE no. 227463), the German Federal Ministry of Education and Research (project no. 01LK1222A), the Swiss National Science Foundation (project nos 200020_135307, 200021_140663, 206021_144947/1 and 20FI20_149002/1), the Academy of Finland (Center of Excellence project no. 1118615), the Academy of Finland (135054, 133872,

251427, 139656, 139995, 137749, 141217, 141451), the Finnish Funding Agency for Technology and Innovation, the Väisälä Foundation, the Nesslering Foundation, the Austrian Science Fund (FWF; project no. L593), the Portuguese Foundation for Science and Technology (project no. CERN/FP/116387/2010), the Swedish Research Council, Vetenskapsrådet (grant 2011-5120), the Presidium of the Russian Academy of Sciences and Russian Foundation for Basic Research (grant 12-02-91522-CERN), the UK Natural Environment Research Council (grant NE/K015966/1), the Royal Society (Wolfson Merit Award), the US National Science Foundation (grants AGS1136479, AGS1447056 and CHE1012293), Caltech ESE Grant (Davidow Foundation), Dreyfus Award EP-11-117, the French National Research Agency (ANR), the Nord-Pas de Calais, and the European Funds for Regional Economic Development (FEDER, Labex-Cappa, ANR-11-LABX-0005-01).

Author Contributions J.A., H.G., A.K., T.N., J.T. and C.W. analysed the nucleation rates; C.Fr. analysed the API-TOF charged clusters; M.H., M.Sim. and C.Y. performed the CI-API-TOF HOM and H₂SO₄ analyses; A.-K.B. analysed the PTR-TOF α -pinene; J.H.S. and X.Z. analysed the ELVOC structures and formation mechanisms; I.K.O. performed the quantum chemical calculations; A.Ad., J.A., A.Am., A.-K.B., F.B., M.B., S.B., J.Cu., J.Cr., A.D., J.Do., J.Du., S.E., C.Fr., C.Fu., H.G., M.H., C.R.H., T.J., H.J., J.Ka., J. Kim, J.Kir., M.Kr., A.K., K.L., V.M., U.M., T.N., F.P., T.P., A.P.P., M.P.R., N.S., K.S., M.Sim., M.Sip., G.S., A.T., J.T., A.W., D.W., R.W., C.W.,

C.Y. and P.Y. collected the data and contributed to the analysis; K.S.C., H.G., K.P., A.R., N.A.D.R., K.S. and C.E.S. evaluated the atmospheric relevance; J.Kir. wrote the manuscript; J.A., J.Do., N.M.D., C.Fr., H.G., M.H., J.H.S., M.Sim., C.W., R.W., C.Y. and X.Z. contributed to Methods and Extended Data; and U.B., K.S.C., J.Cu., J.Do., N.M.D., R.C.F., A.H., J.Kir., M.Ku., J.H.S. and D.R.W. contributed to data interpretation and editing of manuscript. All authors contributed to the development of the CLOUD facility and analysis instruments, and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.Kir. (jasper.kirkby@cern.ch).



This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

METHODS

Overview of the CLOUD facility. The CLOUD experiment at CERN is designed to study the effects of cosmic rays on aerosols, cloud droplets and ice particles, under precisely controlled laboratory conditions. The 3-m-diameter stainless-steel CLOUD chamber and its gas system have been built to the highest technical standards of cleanliness and performance. The CLOUD chamber is periodically cleaned by rinsing the walls with ultra-pure water, followed by heating to 373 K and flushing at a high rate with humidified synthetic air and elevated ozone (several parts per million by volume). Contaminant levels of condensable vapours are in the sub-p.p.t.v. range. The high cleanliness of the chamber, together with its large volume (26.1 m³) and highly stable operating conditions, allows particle formation to be studied under atmospheric conditions at nucleation rates between about 0.001 cm⁻³ s⁻¹ and 100 cm⁻³ s⁻¹. The loss rate of condensable vapours and particles onto the chamber walls is comparable to the ambient condensation sink of the pristine boundary layer.

Ion production in the chamber can be controlled using an internal electric clearing field (which creates an ion-free environment), GCRs or an adjustable π^+ beam^{9,33} from the CERN Proton Synchrotron. The π^+ beam is de-focused to a transverse size of about 1.5 m × 1.5 m when it passes through the CLOUD chamber. With the electric field set to zero, the equilibrium ion-pair concentration in the chamber due to GCRs is around 700 cm⁻³. With the π^+ beam, this can be increased to any value up to about 3,000 cm⁻³. Hence, ion concentrations corresponding to any altitude in the troposphere can be generated in the CLOUD chamber.

The experiment has precise control of the trace vapours inside the chamber and also of the environmental temperature between 300 K and 203 K. Uniform mixing is achieved with magnetically coupled stainless-steel fans mounted at the top and bottom of the chamber. The characteristic gas mixing time in the chamber is a few minutes, depending on the fan speeds. Photochemical processes are initiated by illumination with an ultraviolet fibre-optic system, providing highly stable gas-phase reactions with a precise start time. The contents of the chamber are continuously analysed by instruments connected to sampling probes that project into the chamber. The sampling analysers are tailored for each experimental campaign, but typically comprise around 30–35 instruments, of which up to 10 are mass spectrometers.

Summary of analysing instruments. For the results reported here, the analysing instruments attached to the chamber included a chemical ionization mass spectrometer (CIMS) for H₂SO₄ concentration³⁴; an atmospheric pressure interface time-of-flight (API-TOF; Aerodyne Research Inc. and ToFwerk AG)³⁵ mass spectrometer for molecular composition of positively or negatively charged ions and clusters; two chemical ionization atmospheric pressure interface time-of-flight (CI-API-TOF; Aerodyne Research Inc. and ToFwerk AG)^{36,37} mass spectrometers for molecular composition and concentration of neutral gas-phase H₂SO₄ and HOMs; a proton transfer reaction time-of-flight (PTR-TOF; Ionicon Analytik GmbH)³⁸ mass spectrometer for organic vapours; a neutral cluster and air ion spectrometer (NAIS; Airel Ltd)³⁹ for concentrations of positive ions, negative ions and charged clusters in the range 1–40 nm; a nano-radial differential mobility analyser (nRDMA)⁴⁰ and a nano scanning mobility particle sizer (nano-SMPS) for particle size spectra; and several condensation particle counters (CPCs) with 50% detection efficiency thresholds between 1 nm and 4 nm: two Airmodus A09 particle size magnifiers, PSM⁴¹, (one fixed-threshold and the other scanning), two diethyleneglycol CPCs, DEG-CPC^{42,43}, a butanol TSI 3776 CPC and a water TSI 3786 CPC (TSI Inc.).

Additional gas analysers included dew-point sensors (EdgeTech), sulfur dioxide (Thermo Fisher Scientific, Inc. 42i-TLE) and ozone (Thermo Environmental Instruments TEI 49C). For certain tests, HONO vapour was supplied to the chamber and photolysed with ultraviolet light to produce OH· in the absence of O₃. The gaseous HONO was generated by continual mixing of H₂SO₄ with NaNO₂ (ref. 44) in a specially designed stainless-steel reactor, and then steadily flowed into the chamber. The HONO analyser involved a specially designed probe that passed samples of air from the chamber through a solution of H₂SO₄ and sulfanilamide, which was then analysed online with a long path absorption photometer (LOPAP)⁴⁵.

Determination of the nucleation and growth rates. The nucleation rates (in cm⁻³ s⁻¹) were measured under neutral (J_n), ground-level GCR (J_{GCR}) and π^+ beam (J_π) conditions. Neutral nucleation rates are measured with the clearing field electrodes set to ± 30 kV, which establishes an electric field of about 20 kV m⁻¹ in the chamber. This completely suppresses ion-induced nucleation because, under these conditions, small ions or molecular clusters are swept from the chamber in about 1 s. Because all of the nucleation and growth processes under consideration take place on substantially longer timescales, neutral nucleation rates can be measured with zero background from ion-induced nucleation. For GCR and π^+ beam conditions, the electric field was set to zero, leading to equilibrium ion-pair concentrations around 700 cm⁻³ and 3,000 cm⁻³, respectively. The nucleation rate

J_n measures the neutral rate alone, whereas J_{GCR} and J_π measure the sum of the neutral and ion-induced nucleation rates, $J_n + J_{\text{ion}}$.

The nucleation rates reported here were obtained primarily with the Airmodus scanning PSM at 1.8-nm threshold (PSM1.8) and the TSI 3776 CPC (CPC2.5), nominally 2.5-nm threshold, but measured at 3.2-nm threshold with WO_x particles⁴⁶. The nucleation rates $J_{1.7}$ are determined at 1.7-nm mobility diameter (1.4-nm mass diameter), at which size a particle is normally considered to be above its critical size and, therefore, thermodynamically stable. The critical size corresponds to the cluster size at which the evaporation and growth rates are equal. It varies with temperature, chemical species, charge and vapour concentrations, and may even be absent when evaporation rates are highly suppressed, such as for sulfuric acid–dimethylamine clusters^{10,37}. Our measurements indicate that the smallest neutral HOM clusters are relatively unstable; therefore, 1.7 nm, which is equivalent to around 5 HOM monomer units, is a reasonable size at which to derive the experimental nucleation rates.

AEROCLOUD model. To determine nucleation rates at 1.7 nm, the time-dependent particle concentrations measured with the PSM1.8 and CPC2.5 are fitted with a simplified numerical model (AEROCLOUD) that treats particle nucleation and growth kinetically at the molecular level. The model uses HOM monomer, HOM dimer and H₂SO₄ production rates derived from the CI-API-TOF experimental data. The measured HOM production rates are scaled by a factor of 1.8 to match the observed particle appearance times and growth rates. This scaling results in good agreement of the model with the experimental data over the full experimental range of HOM concentrations. The scaling factor is within the systematic measurement uncertainty of the CI-API-TOF, and could arise if a nitrate CI-API-TOF does not detect all the HOMs that contribute to particle growth.

Primary ions from GCRs are generated in the model at the known rate of $q = 1.7$ ion pairs per cubic centimetre per second. A fixed parameter of the model, f_c , accounts for the charge sign asymmetry due to differences in the diffusional loss rates of positive and negative primary ions to the chamber walls:

$$q_+ = f_c(2q) \\ q_- = (1 - f_c)(2q)$$

The parameter f_c is determined by the experimentally measured positive and negative ion concentrations in the NAIS to have the value 0.52.

Molecules and particles collide kinetically, and cluster with each other. The model uses a reduced clustering probability (termed a 'sticking probability' below) to account for unstable small clusters, rather than allowing clusters to evaporate once they have formed. This greatly increases the speed of the computation. If the particle formed by a collision exceeds a certain size (corresponding to around 1.7-nm mobility diameter for pure biogenic clusters; see below), then it is assumed to be effectively stable and subsequently grows at near the kinetic limit. The particle growth rate between the PSM1.8 and CPC2.5 is therefore implicitly treated in the model essentially as kinetically limited growth by particle coagulation plus HOM and H₂SO₄ vapour condensation. Particles grow through size bins that are linearly spaced for small sizes and logarithmically spaced from about 2 nm to a maximum size of 400 nm. The time-steps for clustering processes range from 0.9 s to 10 s, depending on the conditions of the experimental run under analysis. The time-step is 10 s for all other processes (for example, updates of gas concentrations, high-voltage clearing-field changes, fan changes, and particle losses due to dilution of the chamber contents or diffusion to the walls). The density of the pure HOM clusters is fixed at 1.3 g cm⁻³, and at 1.85 g cm⁻³ for a pure H₂SO₄ cluster.

For neutral–neutral collisions, the number of particles in size bins 1 and 2 that coagulate in a time interval Δt to produce a particle of mass m_{12} is:

$$n_{12} = K_{00} S'_{00} n_1 n_2 V_{12} \Delta t \quad (1)$$

where K_{00} is the neutral–neutral collision kernel, n_1 , n_2 and n_{12} are the particle number concentrations, and V_{12} is the van der Waals enhancement factor (see below). The neutral–neutral sticking probability for pure biogenic particles, $S'_{00,B}$, is:

$$S'_{00,B} = \exp[-0.693(C_B/m_{12})^{S_B}]$$

where C_B and S_B are free parameters. The parameter C_B effectively defines the threshold mass of stable clusters because the sticking probability $S'_{00,B} = 0.5$ when $C_B = m_{12}$, whereas the parameter S_B controls the sharpness of the threshold. The sticking probability for collisions where at least one particle is mainly sulfuric acid is similarly defined as:

$$S'_{00,A} = \exp[-0.693(C_A/m_{12})^{S_A}]$$

where C_A and S_A are free parameters.

The neutral–neutral collision kernel, K_{00} , in equation (1) is the Fuchs form of the Brownian coagulation coefficient^{47,48}. The van der Waals enhancement factor is the modification to Fuchs theory due to Scaats⁴⁹, as described in ref. 50, for a Knudsen number in the kinetic (free molecular) regime. The enhancement factor is:

$$V_{12} = 1 + \frac{\sqrt{A'/3}}{1 + b_0\sqrt{A'}} + b_1\ln(1 + A') + b_2\ln(1 + A')^3$$

where the reduced Hamaker constant, A' , is:

$$A' = \frac{A}{kT} \frac{r_1 r_2}{(r_1 + r_2)^2}$$

where $r_{1,2}$ are the particle radii, $A = 6.4 \times 10^{-20}$ J (the Hamaker constant for sulfuric acid⁵⁰), $b_0 = 0.0151$, $b_1 = -0.186$, $b_2 = -0.0163$, k is the Boltzmann constant and T is temperature. The same Hamaker constant is used for both sulfuric acid and HOMs because it does not noticeably change the model predictions.

Ions and charged clusters collide according to a similar expression as equation (1):

$$n_{12} = (E \times K_{00}) S' n_1 n_2 \Delta t \quad (2)$$

where E is an enhancement factor to obtain the charged collision kernels (described below). The sticking probability for collisions between a neutral particle and a charged particle, $S'_{0+,0-}$, is:

$$S'_{0+,0-} = \exp[-0.693(C/m_{12})^{S'_{0+,0-}}]$$

where $S'_{0+,0-}$ is a free parameter and $C = C_B$ or C_A for biogenic or acid particles, respectively. Ion–ion recombination results in a neutral particle, which may evaporate at small sizes. The model allows partial evaporation of such recombination particles; in this case the cluster divides into monomers and the mass is conserved. The probability of cluster survival after ion–ion recombination, S'_{+-} , is:

$$S'_{+-} = \exp[-0.693(C_{+-}/m_{12})^{S'_{+-}}]$$

where C_{+-} is a free parameter. A power of unity ($S'_{+-} = 1$) is used because the data do not constrain this parameter well.

To obtain the charged collision kernels, the neutral–neutral collision kernel is multiplied by size-dependent enhancement factors, E :

$$\begin{aligned} E'_{0+,0-} &= K_{0+,0-}/K_{00} \\ E_{+,+-} &= K_{+,+-}/K_{00} \\ E_{+-} &= K_{+-}/K_{00} \end{aligned}$$

where K are the collision kernels and the subscripts refer to the charge of the colliding particles. The charged collision kernels in equation (2) are obtained from ref. 51, which refers to sulfuric acid particles. Because biogenic particles may have different neutral–charged collision kernels, their enhancement factor is left free in the fit:

$$E_{0+,0-} = \frac{E'_{0+,0-} - 1}{f_{0+,0-}} + 1 \quad (3)$$

where $f_{0+,0-}$ is a free parameter.

Ions, monomers, clusters and larger particles are continually lost by diffusion to the walls and by dilution of the chamber contents with fresh gas mixture. The dilution lifetime is near 3 h (10^{-4} s⁻¹), depending on the total sampling rate of all instruments attached to the chamber. The wall loss rate is 1.8×10^{-3} s⁻¹ for H₂SO₄ monomers, and decreases with increasing cluster or molecule diameter as $1/d$. The same scaling law is used to obtain the wall loss rate for HOMs; that is, it is assumed that HOMs and particles that collide with the walls are irreversibly lost. For experimental runs for which there is a pre-existing population of particles in the chamber at the start of a run due to incomplete cleaning of the chamber, losses to this coagulation sink are accounted for by inserting the initial size distribution into the size bins of the model.

To determine the nucleation rates, the five free parameters of the model (S_B , S_A , $S_{0+,0-}$, $f_{0+,0-}$ and C_{+-}) are fitted to the experimental particle concentrations in the PSM1.8 and CPC2.5 versus time. For example, for neutral pure biogenic runs, only one free parameter (S_B) is involved in the fit. The value of S_B ranges from 12 to 14, S_A from 4 to 6, $S_{0+,0-}$ from 0.1 to 1.0, $f_{0+,0-}$ is near 4 and C_{+-} is near 10,000 Th. The parameters C_B , C_A , S_{+-} and f_c were determined by a global fit to all runs in the dataset and then subsequently fixed at these values. The fitted threshold masses for C_B and C_A are around 1,300 Th and 700 Th, respectively. The parameter S_{+-} is set to 1.0 and f_c is set to 0.52. The time development of the particle number

concentrations in both counters throughout all of the nucleation events in our dataset is well reproduced by the model (an example is shown in Extended Data Fig. 4b).

After fitting the data with the model, the nucleation rate $J_{1.7}$ is determined as the number of particles that grow to a mobility diameter of 1.7 nm or larger in any time-step, divided by the time increment. In each nucleation run at fixed conditions, the time t_{\max} is determined at which $J_{1.7}$ is maximum; the value of $J_{1.7}$ for that run is then calculated as the mean measurement over the interval ($t_{\max} \pm 300$ s).

There are three major advantages of using a data-driven kinetic model to determine nucleation rates rather than making direct measurements with the PSM1.8 or CPC2.5 data. First, it avoids the need for time derivatives of the data, which are subject to large errors at low counting rates. Second, particle growth rates are determined by kinetics and properly account for growth due to collisions both with monomers and with other particles. The model treatment of the data therefore avoids the exponential sensitivity on experimental growth rates that occurs with other methods^{52–55}. Experimental growth rates are determined from particle counter rise times and have relatively large uncertainties in the 1–3-nm size range. Finally, the model requires consistency between the PSM1.8 and CPC2.5 so the formation rates are experimentally constrained both near the 1.7-nm threshold size and near 3 nm.

Verification of the model nucleation rates. We performed extensive cross-checks of the nucleation rates obtained with the model by calculating the nucleation rates independently in two additional ways: (1) direct measurements at 1.8 nm using the scanning PSM and (2) CPC2.5 measurements that are stepwise-corrected to 1.7-nm threshold size. Within their experimental uncertainties, the nucleation rates obtained by both these methods agree well with the values obtained with the AEROCLOUD kinetic model.

The stepwise-corrected method is described in detail in ref. 55, but a brief summary is provided here. The nucleation rates are derived from the rate of change of the formation rates, dN_{CPC}/dt , where N_{CPC} is the particle number concentration measured with the CPC2.5 above its detection threshold, d_{th} . The formation rate is corrected in two sequential steps for particle losses to chamber walls, dilution and coagulation: (1) particle losses above d_{th} and (2) particle losses during growth from 1.7 nm to d_{th} . The dilution and wall loss rates are the same as in the kinetic model. To calculate the coagulation rate, the particles are divided into size bins and then the loss rate in each bin i is computed by summing the size-dependent collision (coagulation-loss) rate of the particles in bin i with those in all other bins. The total coagulation loss rate is then the sum of the particle loss rates in each bin i .

Correcting for particle losses during growth from 1.7 nm to d_{th} (item (2) above) requires knowledge of the particle growth rate. This is experimentally determined with several instruments, for example, from the appearance times measured in the scanning PSM⁵⁶, which detects particles over a range of threshold diameters between 1 nm and 2.5 nm. The growth rates were also measured over different size ranges with several other instruments, including a fixed-threshold PSM, two DEGCPCs, a TSI 3776 CPC, an API-TOF, an NAIS, an nRDMA and a nano-SMPS. The experimental growth rates are parameterized because they cannot be measured sufficiently precisely at each point in time during all events. To determine the nucleation rate at 1.7 nm from the corrected formation rate at d_{th} , the size interval is divided into m log-normally spaced bins, $d\log(D_p)$, chosen to match the spacing of the SMPS bins at larger sizes. The residence time of a particle in each bin is $\delta t = \delta d_i / (\text{growth rate})$, where δd_i is the size of bin i . Starting with the measured particle distribution above d_{th} , the size distribution and formation rate is then extended towards 1.7 nm in a stepwise process. In the first step, using the known loss rates due to the chamber walls, dilution and coagulation, as well as the time δt , the concentration in the largest new bin is calculated, as well as the formation rate into this bin. Using this concentration, the size distribution is updated and the process is repeated until, after m steps, the smallest size bin at 1.7 nm is reached, where the nucleation rate is determined.

The NAIS. The neutral cluster and air ion spectrometer (NAIS)⁵⁷ measures the size distributions of positively and negatively charged particles, and also of total (charged plus neutral) particles, between mobility-equivalent diameters of 0.75 nm and 45 nm. Because the instrument includes two mobility analysers operating in parallel, positive and negative spectra are obtained simultaneously, each with 21 electrometers. Taking into account the internal diffusion losses, the mobility distribution is then calculated in 28 size bins from the measured electrometer currents.

The instrument operates sequentially in three modes: ion, particle and offset mode (one cycle takes 150 s). The aerosol sample first passes through a pre-conditioning section containing a discharger, an electric filter, a charger and a second electric filter (post-filter). The charger and discharger are corona needles of opposite polarities. In ion mode, the preconditioning unit is switched off and the sample passes through unaffected. In this way, the mobility analysers

measure only ions and charged particles from the CLOUD chamber. In particle mode—which was not used for the results reported here—both chargers are switched on and so neutral particles from the CLOUD chamber can be classified. The post-filters improve the measurements by removing residual ions from the charger. In offset mode, the dischargers and corresponding filters are switched on. The sample is charged to the opposite polarity as the subsequent analyser and so no detectable particles can enter. In this way, the noise levels and possible parasitic currents are measured to provide corrections for the preceding ion and particle measurement.

After preconditioning, the aerosol sample is classified in two cylindrical mobility analysers. The central electrode consists of several sections, each at a different fixed electric potential. The particles enter the analysers through a circular slit near the central electrode and are collected at the 21 outer electrodes where they transfer their charge to the connected electrometer and the resulting current is measured. The analysers operate at a sheath flow rate of 60 l min⁻¹. Filtered excess air serves as sheath gas to ensure conditions similar to the sample flow. The data inversion that converts the measured electrometer currents to particle concentrations is based on model calculations simulating trajectories of particles with different mobilities, and on calibration measurements of the internal losses. The performance of the NAIS for ion-mobility (size) and concentration measurements is described in refs 58, 59.

The API-TOF mass spectrometer. The atmospheric pressure interface time-of-flight (API-TOF) mass spectrometer¹⁴ measures the mass-to-charge ratio of positive or negative ions with an inlet at atmospheric pressure. The first stage of the instrument consists of an atmospheric pressure interface (API) section where ions are focused and guided by two quadrupoles and an ion lens through three chambers at progressively lower pressures down to 10⁻⁴ mbar. The second stage of the instrument is a time-of-flight (TOF) mass spectrometer at 10⁻⁶ mbar.

The API-TOF was connected to the CLOUD chamber via a 1" (21.7-mm inner diameter) sampling probe shared with the NAIS. A Y-splitter divided the total flow of 20 l min⁻¹ equally between the two instruments. The sample flow for the API-TOF was 0.8 l min⁻¹, with the remainder being discarded.

The API-TOF measurements were made during GCR and π^+ beam runs; that is, the ions were charged by GCRs or charged pions traversing the CLOUD chamber. Because the API-TOF can measure only one polarity at a time, positive and negative ions were measured in different runs. Different instrument settings were used during the campaigns to optimize detection in the low- or high-mass regions of the spectrum. The data were analysed with *tofTools*³⁵, developed by the University of Helsinki. The tool is implemented in MATLAB and allows complete processing of TOF data: averaging, mass calibration, baseline detection, peak fitting and high-resolution analysis.

The CI-API-TOF mass spectrometer. Two nitrate chemical ionization atmospheric pressure interface time-of-flight (CI-API-TOF) mass spectrometers were used to measure neutral sulfuric acid and HOMs. The instruments were operated by the University of Frankfurt (UFRA-CI) and the University of Helsinki (UHEL-CI); differences between the two instruments are indicated in this section by adding the UHEL-CI characteristics in parentheses after those of the UFRA-CI. The CI-API-TOF has been described previously^{36,37}. The sample air from the CLOUD chamber was drawn in through a 1/2" stainless steel tube at flow rate of 9 l min⁻¹ (10 l min⁻¹). An electrostatic filter was installed in front of each instrument to remove ions and charged clusters formed in the chamber. The geometry of both ion sources follows the design of ref. 60, but a corona charger³⁴ (X-ray generator) is used for ion generation. Dry air with nitric acid vapour is flushed over the ionizer to generate NO₃⁻ (HNO₃)_{j=0,2} ions. The ions are guided into the sample flow with an electric field, where they react with sulfuric acid and HOMs. The reaction time is approximately 50 ms (200 ms) before the ions enter the API section through a pinhole with a diameter of 350 μ m (300 μ m). The API section consists of three consecutive differentially pumped chambers where the pressure is progressively reduced and the ions are focused by two sets of quadrupoles and an ion lens system. The mass-to-charge ratios, m/z , of the ions that pass through these chambers are measured by a time-of-flight (TOF) mass spectrometer (Tofwerk AG).

The voltage settings in the API-TOF section influence the mass-dependent transmission efficiency. The transmission curves were determined in a series of calibration measurements in which various perfluorinated acid vapours of different m/z were passed into the instrument in sufficient amounts to saturate all the primary ions. In this way, a constant ion signal could be generated at each m/z and so the transmission efficiency could be determined relative to that of the primary ions mass range. The UFRA-CI operated at the same voltage settings for the entire data collection period; the UHEL-CI was operated in a switching mode between two voltage settings optimized for low and high m/z , respectively.

The raw data were analysed with the MATLAB *tofTools* package³⁵. The mass scale is calibrated to an accuracy of better than 10 p.p.m. using a two-parameter fit.

The concentration of sulfuric acid is calculated from the ratio of bisulfate ion counting rates (in s⁻¹) relative to primary ions as follows:

$$[\text{H}_2\text{SO}_4] = C \times \text{SL}_{\text{H}_2\text{SO}_4} \ln \left[1 + \frac{\text{HSO}_4^- + \text{HSO}_4^- \cdot \text{HNO}_3}{\sum_{j=0}^2 \text{NO}_3^- \cdot (\text{HNO}_3)_j} \right]$$

The factor $\text{SL}_{\text{H}_2\text{SO}_4}$ corrects for losses in the sampling line from the CLOUD chamber. The calibration coefficient, C , is determined by connecting the CI-API-TOF to a well-characterized H₂SO₄ generator⁶¹. The value of C depends on the voltage settings in the API-TOF section and was determined to be $6.5 \times 10^9 \text{ cm}^{-3}$ ($1.2 \times 10^{10} \text{ cm}^{-3}$ and $2.8 \times 10^9 \text{ cm}^{-3}$ for the high and low m/z settings, respectively), with an uncertainty of +50%/-33%. The H₂SO₄ detection limit is $5 \times 10^4 \text{ cm}^{-3}$ or slightly lower.

The concentration of a HOM at $m/z = i$ is calculated as follows:

$$[\text{HOM}] = CT_i \times \text{SL}_{E_1/E_2} \ln \left[1 + \frac{\text{HOM}_i \cdot \text{NO}_3^-}{\sum_{j=0}^2 \text{NO}_3^- \cdot (\text{HNO}_3)_j} \right]$$

Here, $\text{HOM}_i \cdot \text{NO}_3^-$ is the background-subtracted counting rate of the HOM. Background levels were measured by sampling air from the clean CLOUD chamber without any α -pinene present. The factor T_i is the mass-dependent transmission efficiency. The calibration coefficient, C , is the same as that obtained for sulfuric acid because HOMs and sulfuric acid were shown to have similar molecular collision rates with the nitrate ions¹⁶. Furthermore, the binding of NO₃⁻ with highly oxidized HOMs is found in the present study to be strong, so clustering should proceed at near the kinetic limit, as it does for NO₃⁻ with sulfuric acid. The factor SL_{E_1/E_2} corrects for losses in the sampling line from the CLOUD chamber. The values were determined for E_1 and E_2 separately, using experimentally determined diffusion coefficients, as $\text{SL}_{E_1} = 1.443$ and $\text{SL}_{E_2} = 1.372$.

The HOM monomers, E_1 , are the background-subtracted sum of the peaks in the m/z band 235–424 Th; the HOM dimers, E_2 , are the corresponding sum for 425–625 Th. Instrumental contamination peaks are excluded from the band summation, as are peaks assigned to the RO₂ radical (C₁₀H₁₅O_{6,8,10,12}, which correspond to $m/z = 293$ Th, 325 Th, 357 Th and 389 Th). Total HOMs is defined as the sum $\text{RO}_2 + E_1 + E_2$.

HOM yields. The HOM yields from either ozonolysis or OH· chemistry were calculated by assuming equal production and loss rates during steady-state¹⁶:

$$\frac{d[\text{HOM}]}{dt} = \gamma_{\text{Ox}} k_{\text{AP}+\text{Ox}} [\text{AP}] [\text{Ox}] - k_{\text{loss}} [\text{HOM}] = 0$$

where the yield, γ_{Ox} , is the fraction of α -pinene (AP) oxidation reactions leading to HOM formation, and 'Ox' signifies O₃ or OH·. The values of the rate constants (in cm³ per molecule per second) at 278 K for oxidation of α -pinene are $k_{\text{AP}+\text{O}_3} = 8.05 \times 10^{-17}$ and $k_{\text{AP}+\text{OH}} = 5.84 \times 10^{-11}$, from the International Union of Pure and Applied Chemistry (IUPAC)⁶² (the α -pinene + O₃ rate constant is updated on the IUPAC website at http://iupac.pole-ether.fr/htdocs/datasheets/pdf/Ox_VOC8_O3_apinene.pdf). The HOM wall loss rate was determined to be $1.1 \times 10^{-3} \text{ s}^{-1}$, assuming they are irreversibly lost. An additional loss is due to dilution of the chamber contents by makeup gases ($0.1 \times 10^{-3} \text{ s}^{-1}$). The total loss rates for HOMs is then $k_{\text{loss}} = 1.2 \times 10^{-3} \text{ s}^{-1}$.

During the experiments involving pure OH· chemistry, nitrous acid (HONO) concentrations ranging from 0.5 p.p.b.v. to 3 p.p.b.v. were photolysed by ultraviolet radiation from the fibre optic system to produce OH·. This led to a small contamination of NO in the chamber, which may potentially influence the HOM yield. The OH· concentrations in the CLOUD chamber were estimated using the PTR-TOF measurements of the difference of the α -pinene concentrations with no OH· present (ultraviolet off) and OH· present (ultraviolet on at different intensities). The decrease in α -pinene was due to only OH· reactions, because no O₃ was present in the chamber during these experiments. The accuracy for [OH·] is estimated to be $\pm 30\%$ (1 σ) including uncertainties in α -pinene measurements and reaction rate constant, which leads to a systematic scale uncertainty on the HOM production rate, $k_{\text{AP}+\text{OH}} [\text{AP}] [\text{OH} \cdot]$, of $\pm 40\%$ (1 σ). However, run-to-run uncertainties contribute substantially to the overall uncertainty as indicated by the error bars in Extended Data Fig. 2.

The SO₂-CIMS. The SO₂ chemical ionization mass spectrometer (SO₂-CIMS) uses CO₃⁻ primary ions to convert SO₂ to SO₅⁻, which is then measured in a quadrupole mass spectrometer with an API interface (Georgia Tech). The general design of the ion source is shown in ref. 60, but the primary ions are generated with a corona discharge³⁴. The corona needle holder was modified so that CO₂, O₂ and Ar are fed directly over the corona discharge. In this way, direct contact between the N₂ sheath flow and the discharge needle is avoided, which leads to a reduced

contamination by NO_3^- and maximizes the ratio of CO_3^- to NO_3^- . The reaction scheme for the ionization of SO_2 to SO_5^- can be found in ref. 63. The use of a dry N_2 buffer flow in front of the pinhole of the mass spectrometer evaporates associated water molecules from SO_5^- ions, and so sulfur dioxide is detected in the mass spectrum at $m/z = 112$ Th (SO_5^-).

The SO_2 concentration (in p.p.t.v.) is calculated from the ion count rates, $R_{m/z}$, as follows:

$$\text{SO}_2 = C_5 \ln(1 + R_{112}/R_{60})$$

where R_{112} corresponds to the background-corrected ion count rate of SO_5^- and R_{60} is the ion count rate of the primary ion CO_3^- . The calibration factor C_5 was obtained by periodically calibrating the instrument with a SO_2 gas standard (Carbagas AG) during the campaign. During a calibration, the gas standard was diluted with ultraclean humidified air at 38% relative humidity (the same as that supplied to the CLOUD chamber) to achieve a range of different SO_2 mixing ratios between 12 p.p.t.v. and 11 p.p.b.v. The calibration factor was found to be 1.3×10^5 p.p.t.v., with an estimated uncertainty of $\pm 11\%$. The error includes uncertainties in the flow rates during a calibration and in the gas standard concentration, as well as statistical uncertainties. However, we also observed that temperature changes in the experimental hall where the experiments were conducted led to a drift in the SO_5^- background signal when no SO_2 was applied to the CIMS. This effect contributes to the overall uncertainty and mainly affects the measurement at low SO_2 levels (< 100 p.p.t.v.), with lower precision in this concentration range. For example, at 30 p.p.t.v. SO_2 , the estimated uncertainty is $\pm 23\%$, but it becomes progressively smaller with higher SO_2 levels, reaching $\pm 13\%$ above 100 p.p.t.v. SO_2 . The detection limit of the instrument is 15 p.p.t.v. SO_2 .

Experimental errors. To determine $J_{1.7}$, the measured particle concentrations in the PSM1.8 and CPC2.5 versus time are fitted with the AEROCLOUD model (see above). The nucleation rate error, σ_J , has three main components. The dominant error at slow growth rates is due to uncertainties in the PSM1.8 and CPC2.5 detection thresholds for HOM particles⁶⁴. The threshold error components are first determined numerically for each nucleation measurement by performing additional AEROCLOUD fits after shifting the PSM1.8 particle detection threshold by $+0.2/-0.1$ nm and the CPC2.5 threshold by ± 0.4 nm. This provides four fractional $J_{1.7}$ errors which are then averaged for each counter to provide a mean fractional uncertainty, σ_{psm} and σ_{cpc} , respectively. The total error due to detection threshold uncertainties, σ_{thr} , for the combined fit to the PSM1.8 and CPC2.5 data is then:

$$\frac{1}{\sigma_{\text{thr}}^2} = \frac{1}{\sigma_{\text{psm}}^2} + \frac{1}{\sigma_{\text{cpc}}^2} + \frac{\sigma_{\text{psm}}\sigma_{\text{cpc}}}{(\sigma_{\text{psm}}^2 + \sigma_{\text{cpc}}^2)^{1/2}}$$

The total fractional $J_{1.7}$ error, σ_J , is then obtained by adding σ_{thr} in quadrature with an experimental error due to run-to-run reproducibility under nominally identical chamber conditions, σ_{exp} , and an error to account for model approximations, σ_{model} :

$$\sigma_J^2 = \sigma_{\text{thr}}^2 + \sigma_{\text{exp}}^2 + \sigma_{\text{model}}^2$$

where $\sigma_{\text{exp}} = 30\%$ and $\sigma_{\text{model}} = 50\%$.

The concentration of O_3 is measured with a calibrated instrument and is known to $\pm 10\%$. The α -pinene concentration in the PTR-TOF is known to $\pm 10\%$. As discussed above, the uncertainty on SO_2 is $\pm 13\%$ above 150 p.p.t.v., increasing at lower values to $\pm 23\%$ at 30 p.p.t.v.

For CI-API-TOF measurements, the run-to-run experimental uncertainties are $\pm 10\%$ for $[\text{H}_2\text{SO}_4]$ and $\pm 20\%$ for $[\text{HOM}]$. However, there is a larger overall systematic error that scales all measurements by the same amount. The systematic scale uncertainty for $[\text{H}_2\text{SO}_4]$ is estimated to be $+50\%/-33\%$. This estimate is based on a comparison of $[\text{H}_2\text{SO}_4]$ measurements with a CIMS and a calibrated H_2SO_4 generator⁶¹. The systematic uncertainties for $[\text{HOM}]$ have the following sources and fractional errors (1σ): sulfuric acid calibration (50%), charging efficiency of HOMs in the ion source (25%), mass dependent transmission efficiency (50%) and sampling line losses (20%). This results in an overall systematic scale uncertainty for $[\text{HOM}]$ of $+80\%/-45\%$. The uncertainty in the HOM yield from ozonolysis or hydroxyl chemistry is estimated by adding the $[\text{HOM}]$ uncertainty in quadrature with the errors for α -pinene (10%), O_3 (10%), OH· (30%), HOM wall loss rate (6%) and rate constants (35% for the α -pinene O_3 reaction and 20% for the α -pinene OH· reaction). This results in a mean estimated uncertainty in HOM yield for either ozonolysis or hydroxyl chemistry of $+100\%/-60\%$.

Quantum chemical calculations. To estimate the characteristic binding energies and evaporation rates expected for ELVOC clusters, we chose $\text{C}_{10}\text{H}_{14}\text{O}_7$ (molecular

weight of 246) to represent the ELVOC monomer, E_1 , and $\text{C}_{20}\text{H}_{30}\text{O}_{14}$ (molecular weight of 494) to represent the covalently bound ELVOC dimer, E_2 . Their formation mechanism and structures are shown in Extended Data Figs 3 and 7. To evaluate the effect of charge on the formation of ELVOC clusters, we studied initial molecular clusters of E_1 and E_2 that are either neutral or else include an ion of the type E_1^- , HSO_4^- , NO_3^- or NH_4^+ (Extended Data Table 1).

We calculated formation Gibbs free energies at 278 K, $\Delta G_{278\text{K}}$, of different clusters with the MO62X functional⁶⁵ and the 6-31+G(d) basis set⁶⁶ using the Gaussian09 program⁶⁷. The formation Gibbs free energy can be related to evaporation rate as described in refs 68, 69. In previous works^{10,15}, we used the method proposed in ref. 68 for calculating the formation free energy of different clusters. However, this method is too computationally demanding for the large clusters of the present study. The MO62X functional has been shown to be well suited to the study of atmospheric clusters⁷⁰. Ref. 70 has shown how reducing the basis set from the largest Pople basis set available (6-311++G(3df,3pd)) to the basis set used in this work (6-31+G(d)) leads to differences in the calculated formation free energies below 1 kcal mol^{-1} . Therefore, MO62X/6-31+(d) is a good alternative to the B3RCC2 method⁶⁸ when studying large clusters. We confirmed this by comparing the formation free energies previously calculated¹⁵ using the B3RCC2 method with those calculated here using the MO62X/6-31+G(d) method. The differences were found to be below 2 kcal mol^{-1} .

Parameterization of the pure biogenic nucleation rate. We parameterized the experimentally measured pure biogenic nucleation rates in a form suitable for global aerosol models. The neutral and ion-induced pure biogenic nucleation rates (in $\text{cm}^{-3} \text{ s}^{-1}$) are parameterized as:

$$J_n = a_1[\text{HOM}]^{a_2+a_5/[\text{HOM}]} \quad (4)$$

$$J_{\text{in}} = 2[n_{\pm}]a_3[\text{HOM}]^{a_4+a_5/[\text{HOM}]}$$

where $[n_{\pm}] = [n_+] = [n_-]$ is the small-ion concentration of either sign. Expressions for $[\text{HOM}]$ and $[n_{\pm}]$ are given in equations (7) and (10) below, respectively. The parameters a_n are determined from fits to the data in Fig. 3 and have the values $a_1 = 0.04001$, $a_2 = 1.848$, $a_3 = 0.001366$, $a_4 = 1.566$ and $a_5 = 0.1863$, with $[\text{HOM}]$ expressed in units of 10^7 cm^{-3} . The parameterized rates are shown by the curves in Fig. 3. The R^2 value of the fit is 0.97. The terms a_{1-4} describe simple power laws, whereas the term a_5 accounts for the steepening of the nucleation rate at low HOM concentrations. The nucleation rates are assumed to be independent of temperature, except for the effect of rate constants (equation (6) below), because the experimental measurements exist at only a single temperature.

The HOM concentration in equation (4) is determined from its production and loss rates:

$$\frac{d[\text{HOM}]}{dt} = Y_{\text{HOM}*\text{O}_3}k_{\text{MT}*\text{O}_3}[\text{MT}][\text{O}_3] + Y_{\text{HOM}*\text{OH}}k_{\text{MT}*\text{OH}}[\text{MT}][\text{OH}\cdot] - k_{\text{HOM}}[\text{HOM}] \quad (5)$$

where MT represents total monoterpenes. The IUPAC⁶² reaction rate constants (in cm^3 per molecule per second) for oxidation of α -pinene by ozone and hydroxyl radicals are, respectively:

$$k_{\text{MT}*\text{O}_3} = 8.05 \times 10^{-16} \exp(-640/T) \quad (6)$$

$$k_{\text{MT}*\text{OH}} = 1.2 \times 10^{-11} \exp(440/T)$$

where T (in K) is the temperature (the α -pinene+ O_3 rate constant is updated on the IUPAC website at http://iupac.pole-ether.fr/htdocs/datasheets/pdf/Ox_VOC8_O3_apinene.pdf). The HOM yields in each ozone-monoterpene and hydroxyl-monoterpene reaction are $Y_{\text{HOM}*\text{O}_3}$ and $Y_{\text{HOM}*\text{OH}}$, respectively. The parameter k_{HOM} is the HOM loss rate or, equivalently, the atmospheric condensation sink, CS (in s^{-1}). The condensation sink is determined assuming the diffusion characteristics of a typical α -pinene oxidation product (see appendix A1 of ref. 71). Assuming steady-state in equation (5), the HOM concentration becomes:

$$[\text{HOM}] = \frac{Y_{\text{HOM}*\text{O}_3}k_{\text{MT}*\text{O}_3}[\text{MT}][\text{O}_3] + Y_{\text{HOM}*\text{OH}}k_{\text{MT}*\text{OH}}[\text{MT}][\text{OH}\cdot]}{\text{CS}} \quad (7)$$

where the HOM yield from ozonolysis is $Y_{\text{HOM}*\text{O}_3} = 2.9\%$, and from reaction with the hydroxyl radical is $Y_{\text{HOM}*\text{OH}} = 1.2\%$ (Extended Data Fig. 2). The HOM yield from ozonolysis is determined from CLOUD measurements in the presence of a hydroxyl scavenger (0.1% H_2). The HOM yield from reaction with hydroxyl radicals is determined from CLOUD measurements in the absence of ozone, and where photolysed HONO provides the OH· source. Therefore, the experimental measurement of hydroxyl-initiated oxidation is made in the presence of NO_x , as occurs in the atmosphere.

The small-ion concentration in equation (4) is calculated from the steady-state solution of the ion balance equation:

$$\frac{d[n_{\pm}]}{dt} = q - \alpha[n_{\pm}]^2 - k_i[n_{\pm}] \quad (8)$$

where q (in $\text{cm}^{-3}\text{s}^{-1}$) is the ion-pair production rate and α is the ion-ion recombination coefficient (in cm^3s^{-1}). The factor of 2 in equation (4) accounts for nucleation from positive and negative ions. For the CLOUD GCR data, $q = 1.7\text{ cm}^{-3}\text{s}^{-1}$. Terrestrial radioactivity such as radon contributes additional ionization in the boundary layer over land masses⁷². The ion loss rate, k_i , is due to the condensation sink, CS, and ion-induced nucleation:

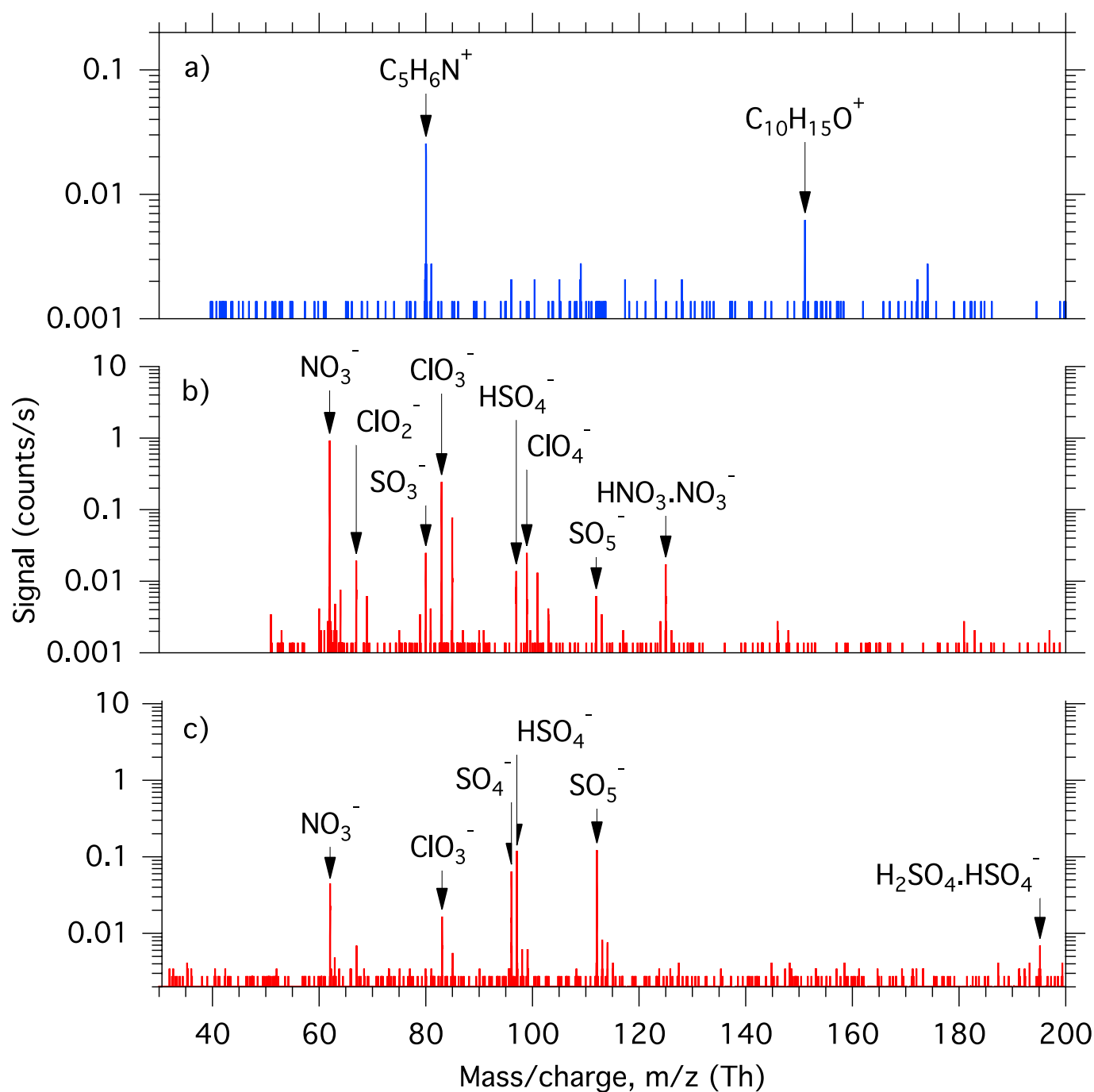
$$k_i = \text{CS} + \frac{J_{\text{in}}}{2[n_{\pm}]} \quad (9)$$

where $J_{\text{in}}/(2[n_{\pm}])$ is given by equation (4) and the steady-state concentration of small ions is, from equation (8):

$$[n_{\pm}] = \frac{(k_i^2 + 4\alpha q)^{1/2} - k_i}{2\alpha} \quad (10)$$

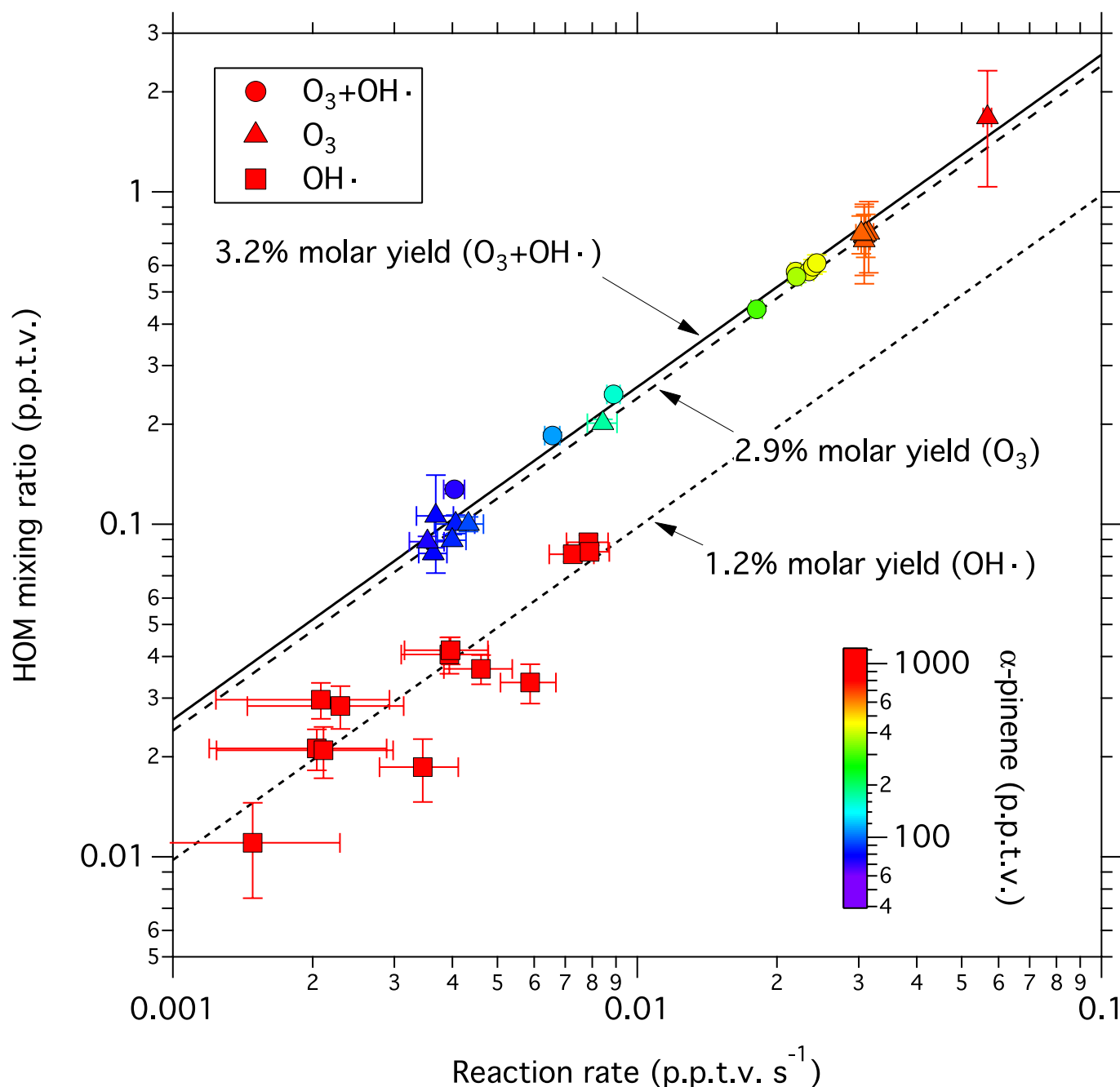
From equations (8) and (9), J_{in} saturates at $2q$ at high nucleation rates (see Fig. 3).

33. Enghoff, M. B., Pedersen, J. O. P., Uggerhøj, U. I., Paling, S. M. & Svensmark, H. Aerosol nucleation induced by a high energy particle beam. *Geophys. Res. Lett.* **38**, L09805 (2011).
34. Kürten, A., Rondo, L., Ehrhart, S. & Curtius, J. Performance of a corona ion source for measurement of sulphuric acid by chemical ionisation mass spectrometry. *Atmos. Meas. Tech.* **4**, 437–443 (2011).
35. Junninen, H. *et al.* A high-resolution mass spectrometer to measure atmospheric ion composition. *Atmos. Meas. Tech.* **3**, 1039–1053 (2010).
36. Jokinen, T. *et al.* Atmospheric sulphuric acid and neutral cluster measurements using CI-API-TOF. *Atmos. Chem. Phys.* **12**, 4117–4125 (2012).
37. Kürten, A. *et al.* Neutral molecular cluster formation of sulfuric acid-dimethylamine observed in real time under atmospheric conditions. *Proc. Natl Acad. Sci. USA* **111**, 15019–15024 (2014).
38. Graus, M., Müller, M. & Hansel, A. High resolution PTR-TOF: quantification and formula confirmation of VOC in real time. *J. Am. Soc. Mass Spectrom.* **21**, 1037–1044 (2010).
39. Mirme, S. *et al.* Atmospheric sub-3 nm particles at high altitude. *Atmos. Chem. Phys.* **10**, 437–451 (2010).
40. Zhang, S. H., Akutsu, Y., Russell, L. M., Flagan, R. C. & Seinfeld, J. H. Radial differential mobility analyzer. *Aerosol Sci. Technol.* **23**, 357–372 (1995).
41. Vanhanen, J. *et al.* Particle size magnifier for nano-CN detection. *Aerosol Sci. Technol.* **45**, 533–542 (2011).
42. Iida, K., Stolzenburg, M. R. & McMurry, P. H. Effect of working fluid on sub-2 nm particle detection with a laminar flow ultrafine condensation particle counter. *Aerosol Sci. Technol.* **43**, 81–96 (2009).
43. Wimmer, D. *et al.* Performance of diethylene glycol-based particle counters in the sub-3 nm size range. *Atmos. Meas. Tech.* **6**, 1793–1804 (2013).
44. Taira, M. & Kanda, Y. Continuous generation system for low-concentration gaseous nitrous acid. *Anal. Chem.* **62**, 630–633 (1990).
45. Heland, J., Kleffmann, J., Kurtenbach, R. & Wiesen, P. A new instrument to measure gaseous nitrous acid (HONO) in the atmosphere. *Environ. Sci. Technol.* **35**, 3207–3212 (2001).
46. Riccobono, F. *et al.* Contribution of sulfuric acid and oxidized organic compounds to particle formation and growth. *Atmos. Chem. Phys.* **12**, 9427–9439 (2012).
47. Fuchs, N. A. *The Mechanics of Aerosols* (Pergamon, 1964).
48. Seinfeld, J. H. & Pandis, S. N. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change* 2nd edn, 600 (Wiley, 2006).
49. Scaets, M. G. Brownian coagulation in aerosols—the role of long range forces. *J. Coll. Interf. Sci.* **129**, 105–112 (1989).
50. Chan, T. W. & Mozurkewich, M. Measurement of the coagulation rate constant for sulfuric acid particles as a function of particle size using tandem differential mobility analysis. *J. Aerosol Sci.* **32**, 321–339 (2001).
51. Laakso, L. *et al.* Kinetic nucleation and ions in boreal forest particle formation events. *Atmos. Chem. Phys.* **4**, 2353–2366 (2004).
52. Kerminen, V.-M. & Kulmala, M. Analytical formulae connecting the “real” and the “apparent” nucleation rate and the nuclei number concentration for atmospheric nucleation events. *J. Aerosol Sci.* **33**, 609–622 (2002).
53. Kulmala, M. & Kerminen, V.-M. On the formation and growth of atmospheric nanoparticles. *Atmos. Res.* **90**, 132–150 (2008).
54. Ehrhart, S. & Curtius, J. Influence of aerosol lifetime on the interpretation of nucleation experiments with respect to the first nucleation theorem. *Atmos. Chem. Phys.* **13**, 11465–11471 (2013).
55. Kürten, A., Williamson, C., Almeida, J., Kirkby, J. & Curtius, J. On the derivation of particle nucleation rates from experimental formation rates. *Atmos. Chem. Phys.* **15**, 4063–4075 (2015).
56. Lehtipalo, K. *et al.* Methods for determining particle size distribution and growth rates between 1 and 3 nm using the Particle Size Magnifier. *Bor. Environ. Res.* **19** (Suppl. B), 215–236 (2014).
57. Mirme, S. & Mirme, A. The mathematical principles and design of the NAIS – a spectrometer for the measurement of cluster ion and nanometer aerosol size distributions. *Atmos. Meas. Tech.* **6**, 1061–1071 (2013).
58. Asmi, E. *et al.* Results of the first air ion spectrometer calibration and intercomparison workshop. *Atmos. Chem. Phys.* **9**, 141–154 (2009).
59. Gagné, S. *et al.* Intercomparison of air ion spectrometers: an evaluation of results in varying conditions. *Atmos. Meas. Tech.* **4**, 805–822 (2011).
60. Eisele, F. L. & Tanner, D. J. Measurement of the gas-phase concentration of H_2SO_4 and methane sulfonic acid and estimates of H_2SO_4 production and loss in the atmosphere. *J. Geophys. Res. Atmos.* **98**, 9001–9010 (1993).
61. Kürten, A., Rondo, L., Ehrhart, S. & Curtius, J. Calibration of a chemical ionization mass spectrometer for the measurement of gaseous sulphuric acid. *J. Phys. Chem. A* **116**, 6375–6386 (2012).
62. Atkinson, R. *et al.* Evaluated kinetic and photochemical data for atmospheric chemistry: volume II – gas phase reactions of organic species. *Atmos. Chem. Phys.* **6**, 3625–4055 (2006).
63. Möhler, O., Reiner, T. H. & Arnold, F. The formation of SO_4^- by gas-phase ion-molecule reactions. *J. Chem. Phys.* **97**, 8233–8239 (1992).
64. Kangasluoma, J. *et al.* Sub-3 nm particle size and composition dependent response of a nano-CPC battery. *Atmos. Meas. Tech.* **7**, 689–700 (2014).
65. Zhao, Y. & Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **120**, 215–241 (2008).
66. Ditchfield, R., Hehre, W. J. & Pople, J. A. Self-consistent molecular-orbital methods. IX. An extended Gaussian-type basis for molecular-orbital studies of organic molecules. *J. Chem. Phys.* **54**, 724–728 (1971).
67. Frisch, M. J. *et al.* Gaussian 09 Revision A.01, http://www.gaussian.com/g_prod/g09.htm (Gaussian, Inc., 2009).
68. Ortega, I. K. *et al.* From quantum chemical formation free energies to evaporation rates. *Atmos. Chem. Phys.* **12**, 225–235 (2012).
69. Ortega, I. K. *et al.* Can highly oxidized organics contribute to atmospheric new particle formation? *J. Phys. Chem. A* **120**, 1452–1458 (2016).
70. Elm, J. & Mikkelsen, K. V. Computational approaches for efficient modelling of small atmospheric clusters. *Chem. Phys. Lett.* **615**, 26–29 (2014).
71. Mann, G. W. *et al.* Intercomparison of modal and sectional aerosol microphysics representations within the same 3-D global chemical transport model. *Atmos. Chem. Phys.* **12**, 4449–4476 (2012).
72. Zhang, K. *et al.* Radon activity in the lower troposphere and its impact on ionization rate: a global estimate using different radon emissions. *Atmos. Chem. Phys.* **11**, 7817–7838 (2011).
73. Rissanen, M. P. *et al.* The formation of highly oxidized multifunctional products in the ozonolysis of cyclohexene. *J. Am. Chem. Soc.* **136**, 15596–15606 (2014).
74. Rissanen, M. P. *et al.* Effects of chemical complexity on the autooxidation mechanisms of endocyclic alkene ozonolysis products: from methylcyclohexenes toward understanding α -pinene. *J. Phys. Chem. A* **119**, 4633–4650 (2015).



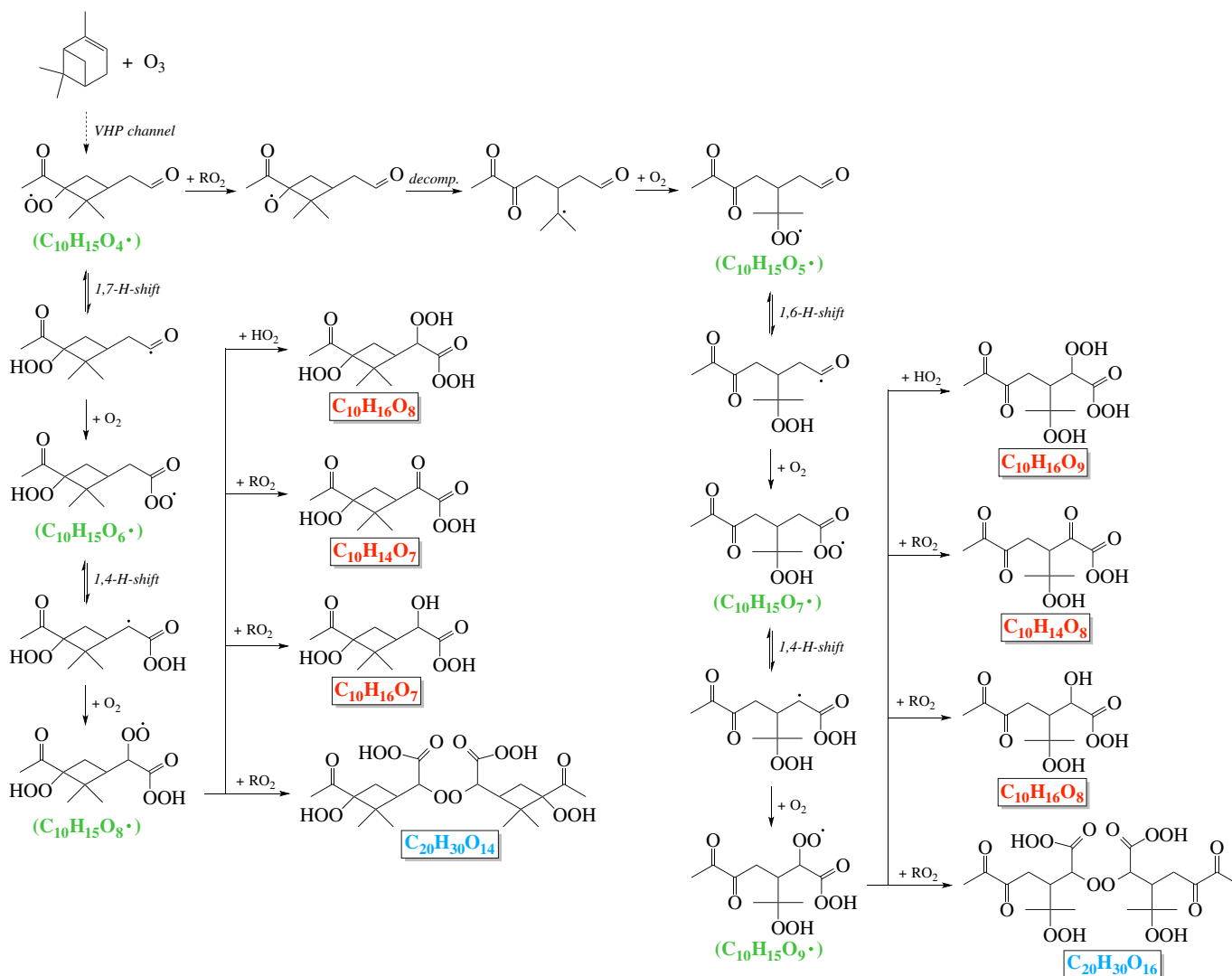
Extended Data Figure 1 | Small-ion mass spectra. **a, b,** Composition of positive (**a**) and negative (**b**) small ions measured by the API-TOF under GCR conditions and before adding any SO_2 to the chamber. The experimental conditions are zero α -pinene, 35 p.p.b.v. O_3 , zero H_2 or HONO, 38% relative humidity, 278 K and $[\text{H}_2\text{SO}_4] < 5 \times 10^4 \text{ cm}^{-3}$. Collisions will transfer positive charge to contaminant molecules having the highest proton affinity (**a**), and negative charge to contaminant molecules with the lowest proton affinity, that is, highest gas-phase acidity (**b**). From molecular cluster measurements, the positive ions

also include ammonium (NH_4^+), but its mass is below the set acceptance cut-off of the API-TOF. **c,** The negative small-ion spectrum at $[\text{H}_2\text{SO}_4] = 1.2 \times 10^5 \text{ cm}^{-3}$, after adding 32 p.p.t.v. SO_2 to the chamber, showing that the dominant ions species shift from nitrate to sulfur-containing. The experimental conditions are 340 p.p.t.v. α -pinene, 35 p.p.b.v. O_3 , zero H_2 or HONO, 38% relative humidity and 278 K. Water molecules evaporate rapidly from most hydrated ions in the API-TOF and so are not detected.



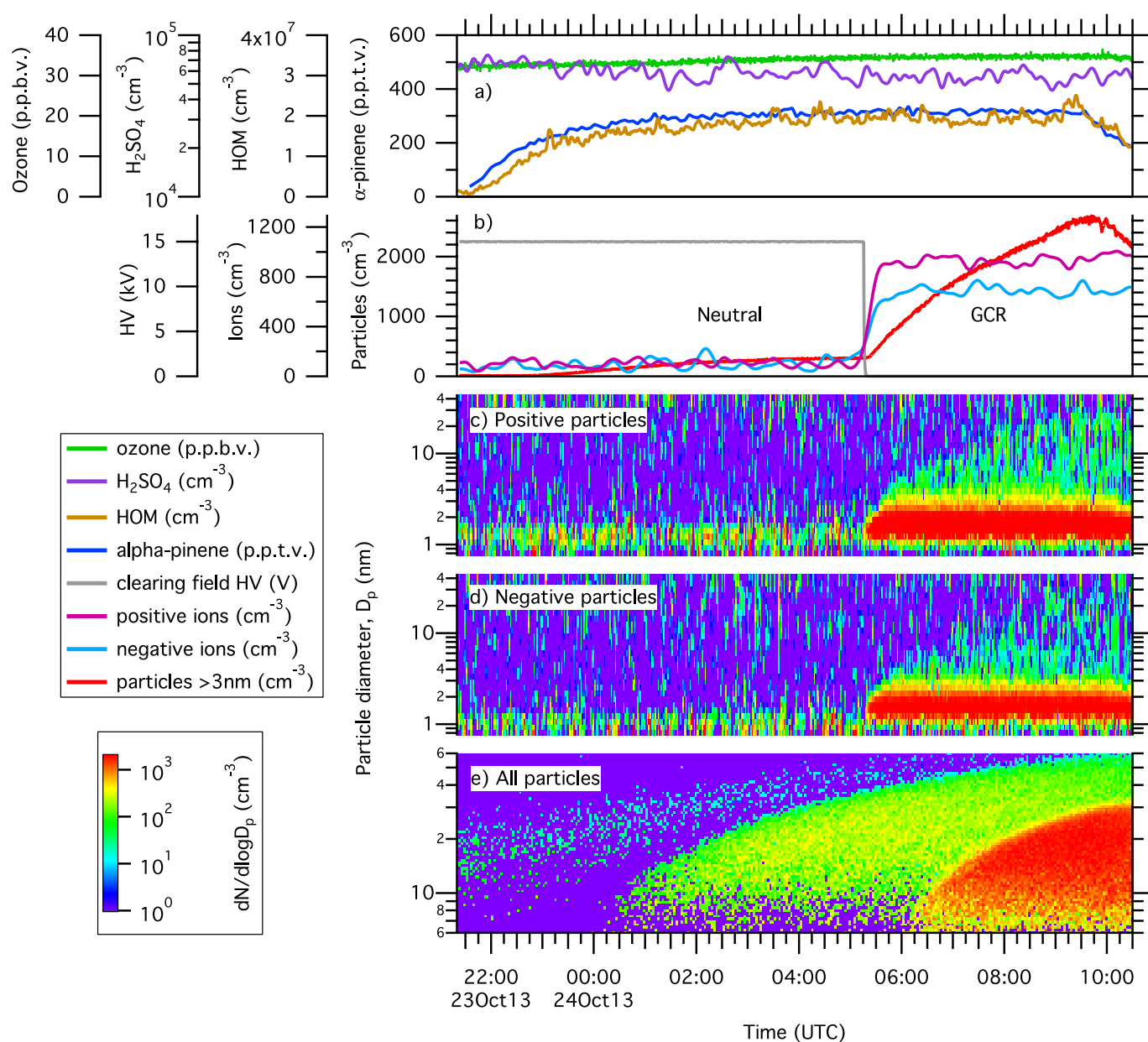
Extended Data Figure 2 | HOM yields versus α -pinene oxidation rates with O₃ and OH. Total HOM mixing ratios versus α -pinene reaction rate with (i) O₃ plus OH· (ozone without H₂ scavenger; circles and solid line), (ii) O₃ alone (ozone with 0.1% H₂ scavenger; triangles and dashed line) and (iii) OH· alone (produced by ultraviolet photolysis of nitrous acid, HONO, in the absence of O₃; squares and dotted line). The yields are shown for total HOMs = RO₂· + E₁ + E₂. The experimental conditions are 38% relative humidity, 278 K and (i) 70–440 p.p.t.v. α -pinene,

21–35 p.p.b.v. O₃, zero H₂ or HONO, 0%–100% ultraviolet, (ii) 80–1,230 p.p.t.v. α -pinene, 21–35 p.p.b.v. O₃, 0.1% H₂, zero HONO, 0%–100% ultraviolet, and (iii) 840–910 p.p.t.v. α -pinene, zero O₃ or H₂, 0.5–3 p.p.b.v. HONO, 0%–100% ultraviolet. The bars indicate 1 σ point-to-point errors. Overall systematic scale uncertainties of $\pm 40\%$ for the reaction rates and $+80\%$ – -45% for the HOM mixing ratios are not shown. The combined errors on the HOM molar yields for either ozonolysis or hydroxyl chemistry are $+100\%$ – -60% ($\pm 1\sigma$).



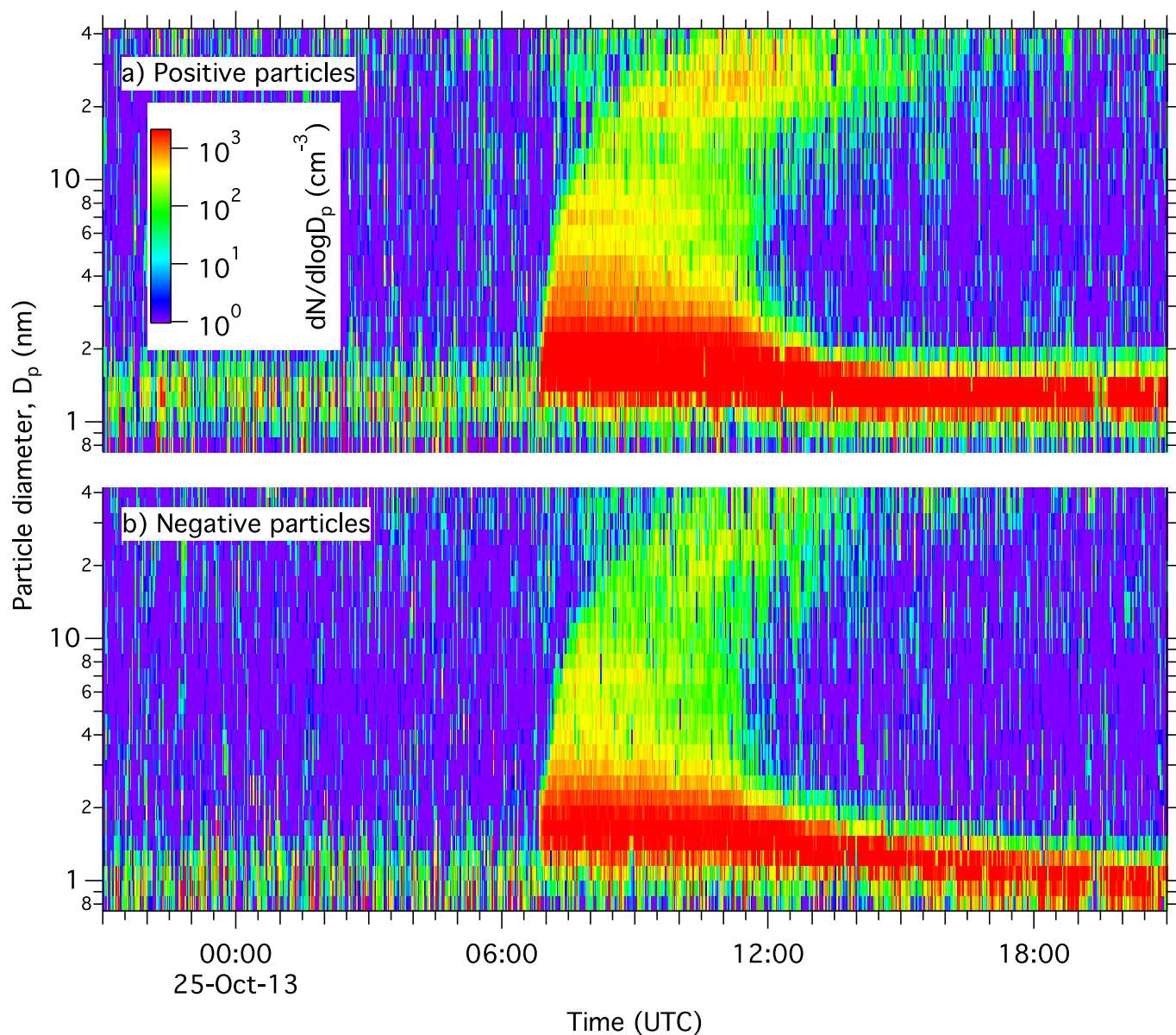
Extended Data Figure 3 | Proposed mechanism for the formation of the E₁ and E₂ surrogates via peroxy radical formation. The proposed scheme for the formation of the ELVOC monomer ($C_{10}H_{14}O_7$) and dimer ($C_{20}H_{30}O_{14}$) surrogates selected for quantum chemical calculations (Extended Data Fig. 7) is based on recently established autooxidation mechanisms for a series of cycloalkane + O_3 systems^{16,18–20,73,74}. Peroxy radicals in the figure are indicated by a green label, E₁ by a red label and E₂ by a blue label. Addition of ozone to the double bond of α -pinene produces two carbonyl-substituted Criegee biradicals. The energy-rich Criegee biradical is either collisionally stabilized, or isomerizes via 1,4-H-shift to a vinylhydroperoxide (VHP), which then decomposes to yield an OH· and an alkenoxy radical. The alkenoxy radical reacts with O_2 , leading to a peroxy radical, which is the potential precursor to a sequence of autooxidation reactions leading to the formation of HOMs¹⁶. Here the peroxy radical $C_{10}H_{15}O_4^\bullet$ is chosen as the starting point for HOM formation. The first intramolecular hydrogen abstraction is likely to take place at the aldehydic carbon from the opposite side of the peroxy group, although the rigid four-carbon-atom ring could hinder bending of the structure. For the *cis* configuration where the peroxy group and the aldehydic hydrogen are on the same side of the cyclobutyl ring, the 1,7-H shift rate is calculated²⁰ to be 0.14 s^{-1} , which initiates the autooxidation chemistry on a fast timescale compared to the HOM lifetime resulting from loss to the CLOUD chamber walls (about 900 s). The resultant acyclic radical undergoes rapid O_2 addition, leading to an -OOH functionalized peroxyacyl radical ($C_{10}H_{15}O_6^\bullet$). The second intramolecular hydrogen abstraction is expected to proceed at the carbon atom in the α position of the peroxyacyl group via 1,4-H isomerization. The resultant $C_{10}H_{15}O_8^\bullet$

terminates by known reactions of peroxy radicals (HO_2^\bullet or RO_2^\bullet under the present experimental conditions), producing a spectrum of HOM monomers that includes the E₁ surrogate, $C_{10}H_{14}O_7$. The homogeneous recombination of two peroxy radicals via elimination of O_2 produces the covalently bound dimer $C_{20}H_{30}O_{14}$ chosen as the E₂ surrogate. Alternatively, $C_{10}H_{15}O_8^\bullet$ can undergo further autooxidation, if sufficiently labile hydrogen atoms are available, leading to the observed closed-shell monomers with $\geq 9\text{ O}$ (Fig. 1). The self/cross-reaction of the $C_{10}H_{15}O_4^\bullet$ peroxy radical produces an alkoxy radical, which decomposes rapidly, leading to a carbonyl-functionalized peroxy radical ($C_{10}H_{15}O_5^\bullet$). This peroxy radical is another potential starting structure for HOM formation. The carbon-ring-opening reaction pathway, while increasing the steric availability of the H atom, might be a slow step. The effective formation rate of the C=O-functionalized peroxy radical is calculated to be less than about 10^{-3} s^{-1} , which is comparable to its wall deposition rate. The timescale with respect to the subsequent autooxidation reaction, on the other hand, is expected to be of the order of seconds, by analogy with that for branched-chain peroxy radicals⁷³. The unbalanced sources and sinks potentially account for the low signals of peroxy radicals with odd oxygen numbers (for example, $C_{10}H_{15}O_5^\bullet$, $C_{10}H_{15}O_7^\bullet$ and $C_{10}H_{15}O_9^\bullet$). The autooxidation process of $C_{10}H_{15}O_5^\bullet$ is presumed to proceed by an autooxidative reaction pathway similar to that for $C_{10}H_{15}O_4^\bullet$, eventually leading to the spectrum of HOM monomers and dimers observed in the CLOUD chamber. Except for the autooxidation channel, all the peroxy radicals are still subject to well-established reactions such as $R(\cdot)O_2 + RO_2/HO_2$, which are potentially important if the reaction rate is comparable to that for the autooxidation channel.



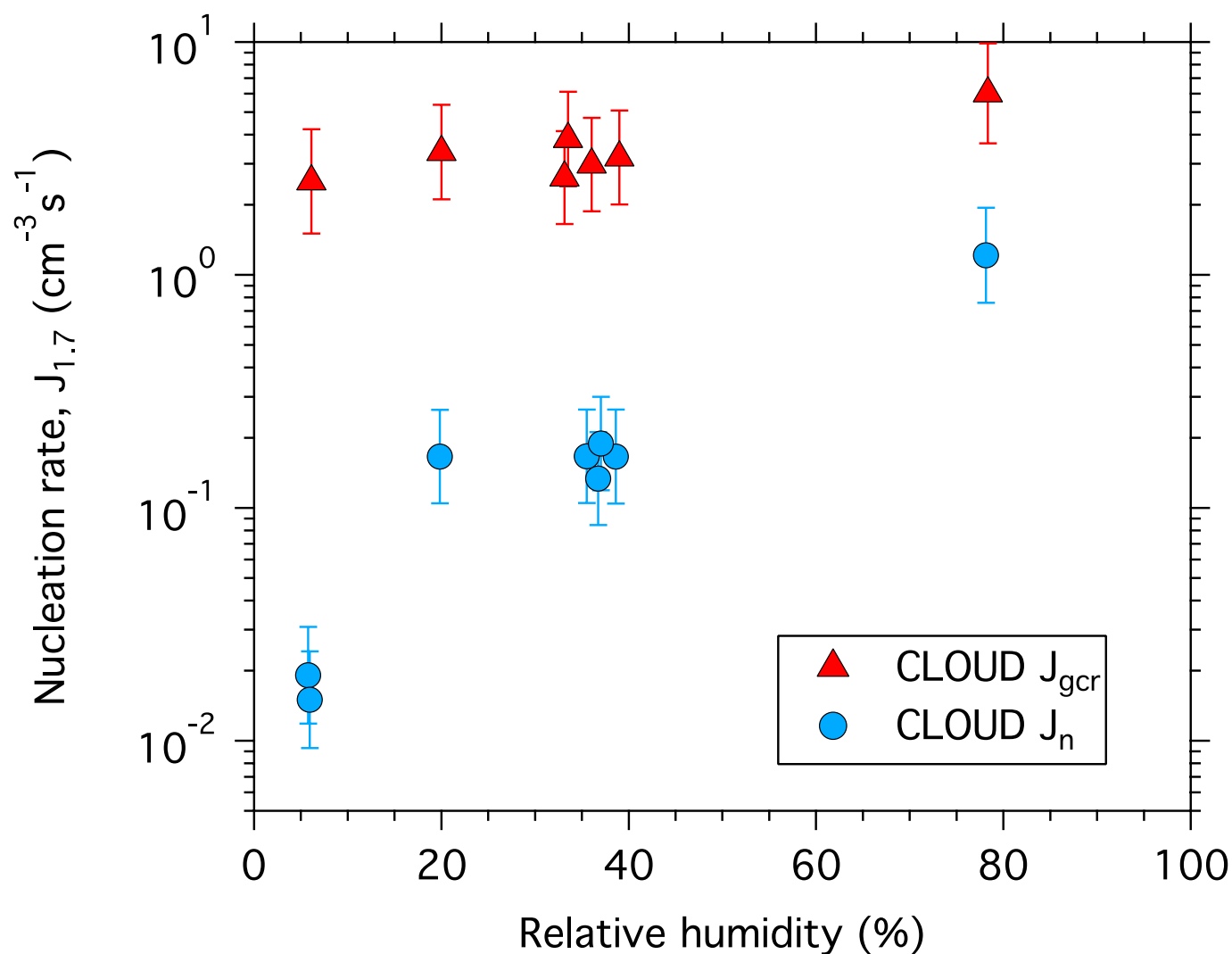
Extended Data Figure 4 | Typical nucleation run sequence. Example of a typical measurement sequence of the neutral and GCR nucleation rates as a function of coordinated universal time (UTC), at zero H_2 or HONO, 38% relative humidity and 278 K. **a**, The run began at 21:22, 23 October 2013, by starting the α -pinene flow into the chamber to reach a chosen equilibrium value near 300 p.p.t.v., which produced an equilibrium total HOMs concentration near $2 \times 10^7 \text{ cm}^{-3}$ (0.8 p.p.t.v.). **b**, Particles (red curve) formed at a slow rate in the chamber without ions present ('neutral' conditions). The clearing field high voltage (HV) was turned off at 05:16, 24 October 2013, and the subsequent presence of ions in the chamber from GCRs caused a sharp increase in the particle formation rate by about one order of magnitude (as seen by the increase in the gradient of the red curve). The nucleation rates are measured under constant gas conditions in the period before ($J_n = 0.14 \text{ cm}^{-3} \text{ s}^{-1}$) and after ($J_{\text{GCR}} = 3.3 \text{ cm}^{-3} \text{ s}^{-1}$) turning off the clearing field high voltage. **c**, **d**, Ion-induced nucleation is

observed both for positive (**c**) and negative (**d**) charged particles, followed by rapid particle growth to sizes above 10 nm. **e**, The nucleated particles grew over a period of several hours to diameters approaching 50 nm, where they begin to constitute cloud condensation nuclei. A sharp increase in the formation rate of particles above the SMPS detection threshold of 5 nm can be seen when GCR ions are present. The colour scale in **c–e** indicates $dN/d\log(D_p)$, where N (in cm^{-3}) is the particle number concentration and D_p (in nm) is the particle diameter. The concentrations of ozone and contaminant H_2SO_4 were essentially constant during the run, which ended at 09:30 when the α -pinene flow to the chamber was turned off. The H_2SO_4 measurement near $5 \times 10^4 \text{ cm}^{-3}$ corresponds to the instrumental background level of the CI-API-TOF mass spectrometer and so represents an upper limit on the actual concentration. Further characteristics of this run can be seen in Fig. 1.



Extended Data Figure 5 | Ion-induced nucleation event without H_2SO_4 , measured in the NAIS. a, b, Example of a nucleation event showing the growth versus time of positive (a) and negative (b) charged particles at 530 p.p.t.v. α -pinene, 35 p.p.b.v. O_3 , zero H_2 or HONO, $3.4 \times 10^7 \text{ cm}^{-3}$ HOM, 38% relative humidity, 278 K and $[\text{H}_2\text{SO}_4] < 5 \times 10^4 \text{ cm}^{-3}$. The colour scale shows the concentration of ions and charged particles. The clearing field high voltage was turned off at 06:48, marking the start of

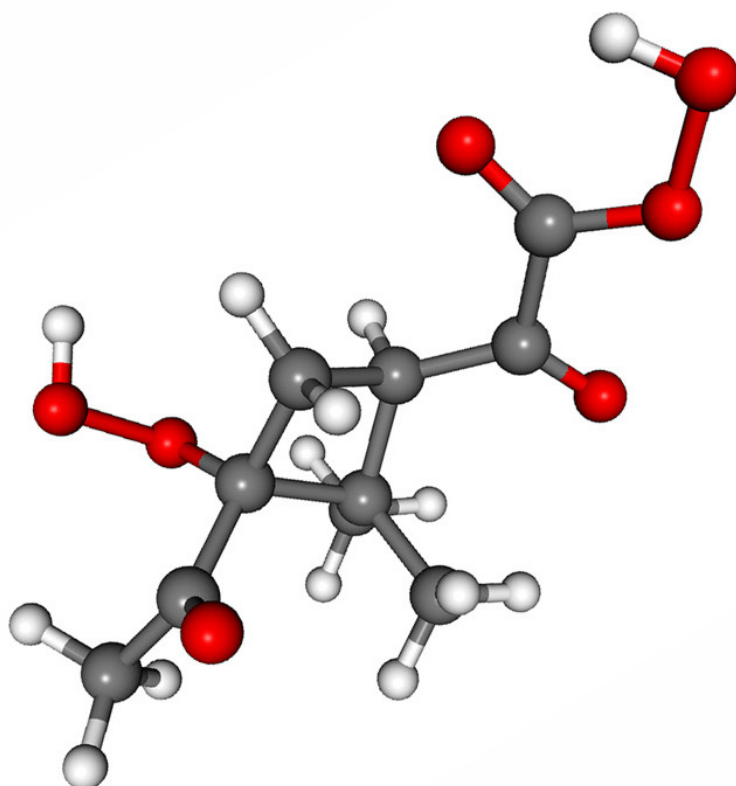
GCR ionization conditions in the chamber, and the α -pinene flow into the chamber was stopped at 10:52. Ion-induced nucleation can be seen for positive and negative charged particles, followed by rapid growth to sizes above 10 nm. Ion-ion recombination progressively neutralizes the charged particles as they grow, but some reappear at larger sizes, owing to diffusion charging.



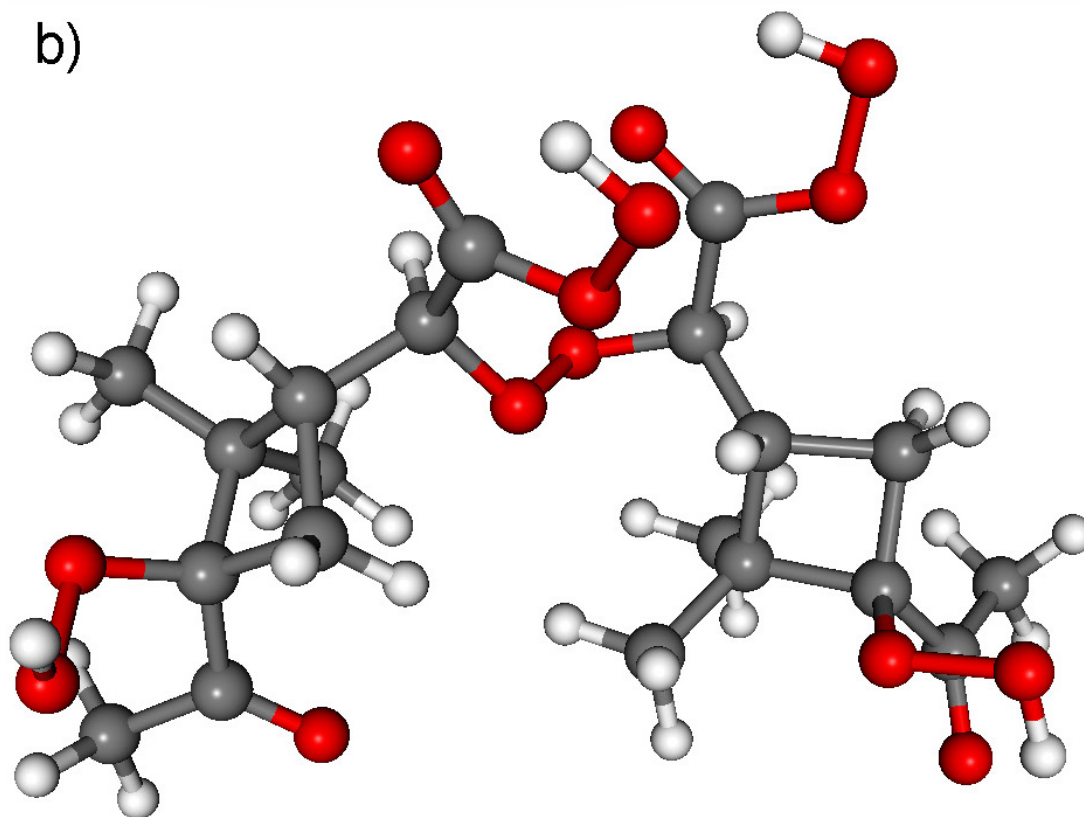
Extended Data Figure 6 | Nucleation rates versus relative humidity. Neutral (J_{n} ; circles) and GCR (J_{gcr} ; triangles) nucleation rates versus relative humidity. The experimental conditions are 250–800 p.p.t.v. α -pinene, 30–35 p.p.b.v. O_3 , zero H_2 or HONO, $(1.1\text{--}2.9) \times 10^7 \text{ cm}^{-3}$ HOM, 278 K and $(0.5\text{--}1.5) \times 10^5 \text{ cm}^{-3}$ H_2SO_4 . All measurements have

been corrected to the same total HOMs concentration ($2.05 \times 10^7 \text{ cm}^{-3}$) using the curves shown in Fig. 3. The bars indicate 1σ total errors, although these are not shown in the x direction because they are smaller than the symbols.

a)

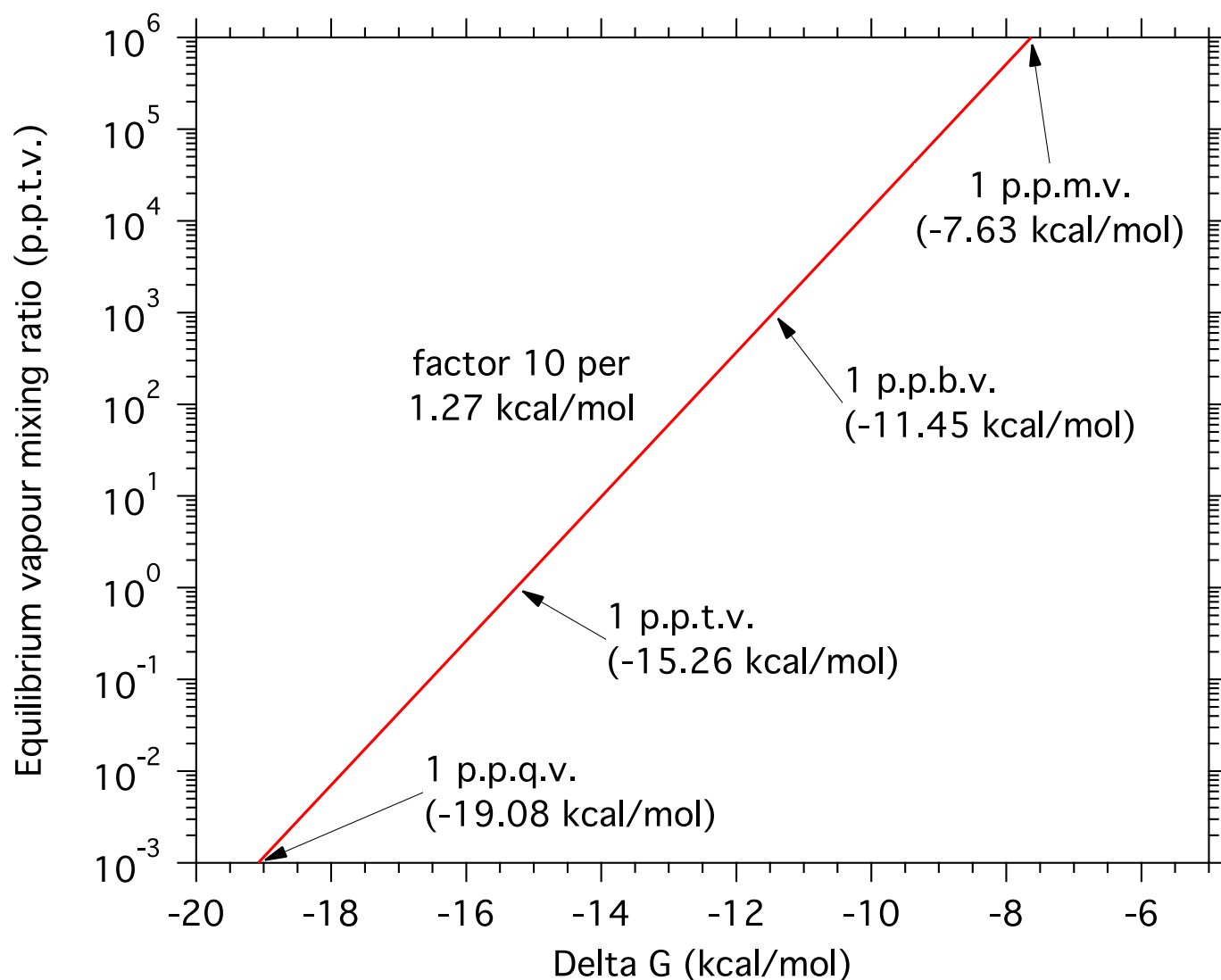


b)



Extended Data Figure 7 | Surrogate molecules chosen for quantum chemical calculations. **a, b,** Structures of the surrogate molecules chosen for quantum chemical calculations to represent the ELVOC monomer, E_1 , $C_{10}H_{14}O_7$ (**a**) and the covalently bound dimer, E_2 , $C_{20}H_{30}O_{14}$ (**b**).

Grey spheres represent carbon atoms, red are oxygen atoms and white are hydrogen atoms. We show their proposed formation mechanisms in Extended Data Fig. 3.



Extended Data Figure 8 | Relationship between cluster formation energies and equilibrium evaporation/condensation rates.

Estimated ELVOC vapour mixing ratios versus the $\Delta G_{278\text{ K}}$ at which the condensation and evaporation rates of the cluster at 278 K are in equilibrium^{68,69}. For example, a formation free energy of $-15.3\text{ kcal mol}^{-1}$

corresponds to equal rates for particle evaporation and vapour condensation at 278 K and 1 p.p.t.v. ELVOC vapour mixing ratio ($2.6 \times 10^7\text{ cm}^{-3}$). The evaporation rate increases by a factor of 10 for each $1.27\text{ kcal mol}^{-1}$ reduction of the cluster formation energy.

Extended Data Table 1 | Quantum chemical calculations of ELVOC cluster formation energies

Neutral clusters		Negative clusters		Positive clusters	
Cluster process	ΔG_{278K} (kcal/mol)	Cluster process	ΔG_{278K} (kcal/mol)	Cluster process	ΔG_{278K} (kcal/mol)
$E_1 + E_1$	-5.76	$E_1 + E_1^-$	-20.95	$E_1 + \text{NH}_4^+$	-22.46
$E_2 + E_1$	-2.15	$E_2 + E_1^-$	-19.90	$E_2 + \text{NH}_4^+$	-30.87
				$E_1 + E_1.\text{NH}_4^+$	-11.71
				$E_2 + E_1.\text{NH}_4^+$	-24.35
				$E_1 + E_2.\text{NH}_4^+$	-15.94
$E_1 + \text{H}_2\text{SO}_4$	-9.90	$E_1 + \text{HSO}_4^-$	-26.97		
$E_2 + \text{H}_2\text{SO}_4$	-12.04	$E_2 + \text{HSO}_4^-$	-30.30		
$E_1 + E_1.\text{H}_2\text{SO}_4$	+2.49	$E_1 + E_1.\text{HSO}_4^-$	-15.28		
$E_2 + E_1.\text{H}_2\text{SO}_4$	+3.13				
$E_1 + E_2.\text{H}_2\text{SO}_4$	-5.69				
		$E_1 + \text{NO}_3^-$	-25.99		
		$E_2 + \text{NO}_3^-$	-25.65		
		$E_1 + E_1.\text{NO}_3^-$	-10.09		

Formation Gibbs free energies at 278 K, ΔG_{278K} , for neutral, negatively charged and positively charged ELVOC clusters. The cluster processes indicate the incident E_1/E_2 vapour molecule + the target cluster. Quantum chemical calculations made at other temperatures (not shown) indicate that the binding energies strengthen by $-1.0 \text{ kcal mol}^{-1}$ per 20 K reduction in temperature. The uncertainty in the calculated energies is less than 2 kcal mol^{-1} . Our calculations indicate the following approximate order for different functional groups to contribute to the cluster binding energies involving HSO_4^- or H_2SO_4 (starting with the strongest): (i) carboxylic acids, R-C(=O)-OH ; (ii) hydroxyls, R-OH ; (iii) hydroperoxy acids, R-C(=O)-O-OH ; (iv) hydroperoxides, R-O-OH ; and (v) carbonyls, R-(R')-C=O . In the case of NH_4^+ , the main interacting group is carbonyl, independently of which other groups are attached to it; therefore ammonium will form stronger clusters with carboxylic acids, hydroperoxy acids or carbonyls than it will with hydroxyls or hydroperoxides.

The role of low-volatility organic compounds in initial particle growth in the atmosphere

Jasmin Tröstl¹, Wayne K. Chuang², Hamish Gordon³, Martin Heinritzi⁴, Chao Yan⁵, Ugo Molteni¹, Lars Ahlm⁶, Carla Frege¹, Federico Bianchi^{1,5,7}, Robert Wagner⁵, Mario Simon⁴, Katrianne Lehtipalo^{1,5}, Christina Williamson^{4,8,†}, Jill S. Craven⁹, Jonathan Duplissy^{5,10}, Alexey Adamov⁵, Joao Almeida³, Anne-Kathrin Bernhammer^{11,12}, Martin Breitenlechner^{11,12}, Sophia Brilke⁴, António Dias³, Sebastian Ehrhart³, Richard C. Flagan⁹, Alessandro Franchin⁵, Claudia Fuchs¹, Roberto Guida³, Martin Gysel¹, Armin Hansel^{11,12}, Christopher R. Hoyle^{1,13}, Tuija Jokinen⁵, Heikki Junninen⁵, Juha Kangasluoma⁵, Helmi Keskinen^{5,14,†}, Jaeseok Kim^{14,†}, Manuel Krapf¹, Andreas Kürten⁴, Ari Laaksonen^{14,15}, Michael Lawler^{14,16}, Markus Leiminger⁴, Serge Mathot³, Ottmar Möhler¹⁷, Tuomo Nieminen^{5,10}, Antti Onnela³, Tuukka Petäjä⁵, Felix M. Piel⁴, Pasi Miettinen¹⁴, Matti P. Rissanen⁵, Linda Rondo⁴, Nina Sarnela⁵, Siegfried Schobesberger^{5,†}, Kamalika Sengupta¹⁸, Mikko Sipilä⁵, James N. Smith^{14,19}, Gerhard Steiner^{5,11,20}, António Tomé²¹, Annele Virtanen¹⁴, Andrea C. Wagner⁴, Ernest Weingartner^{1,†}, Daniela Wimmer^{4,5}, Paul M. Winkler²⁰, Penglin Ye², Kenneth S. Carslaw¹⁸, Joachim Curtius⁴, Josef Dommen¹, Jasper Kirkby^{3,4}, Markku Kulmala⁵, Ilona Riipinen⁶, Douglas R. Worsnop^{5,10,22}, Neil M. Donahue^{2,5} & Urs Baltensperger¹

About half of present-day cloud condensation nuclei originate from atmospheric nucleation, frequently appearing as a burst of new particles near midday¹. Atmospheric observations show that the growth rate of new particles often accelerates when the diameter of the particles is between one and ten nanometres^{2,3}. In this critical size range, new particles are most likely to be lost by coagulation with pre-existing particles⁴, thereby failing to form new cloud condensation nuclei that are typically 50 to 100 nanometres across. Sulfuric acid vapour is often involved in nucleation but is too scarce to explain most subsequent growth^{5,6}, leaving organic vapours as the most plausible alternative, at least in the planetary boundary layer^{7–10}. Although recent studies^{11–13} predict that low-volatility organic vapours contribute during initial growth, direct evidence has been lacking. The accelerating growth may result from increased photolytic production of condensable organic species in the afternoon², and the presence of a possible Kelvin (curvature) effect, which inhibits organic vapour condensation on the smallest particles (the nano-Köhler theory)^{2,14}, has so far remained ambiguous. Here we present experiments performed in a large chamber under atmospheric conditions that investigate the role of organic vapours in the initial growth of nucleated organic particles in the absence of inorganic acids and bases such as sulfuric acid or ammonia and amines, respectively. Using data from the same set of experiments, it has been shown¹⁵ that organic vapours alone can drive nucleation. We focus on the growth of nucleated particles and find that the organic vapours that drive initial growth have extremely low volatilities (saturation concentration less than $10^{-4.5}$ micrograms per cubic metre). As the particles increase in size and the Kelvin barrier falls, subsequent growth is primarily

due to more abundant organic vapours of slightly higher volatility (saturation concentrations of $10^{-4.5}$ to $10^{-0.5}$ micrograms per cubic metre). We present a particle growth model that quantitatively reproduces our measurements. Furthermore, we implement a parameterization of the first steps of growth in a global aerosol model and find that concentrations of atmospheric cloud condensation nuclei can change substantially in response, that is, by up to 50 per cent in comparison with previously assumed growth rate parameterizations.

Two measurement campaigns at the CERN CLOUD (Cosmics Leaving Outdoor Droplets) chamber (Methods) focused on aerosol growth with different levels of sulfuric acid and α -pinene oxidation products. With the chamber at 278 K and 38% relative humidity, tropospheric concentrations of α -pinene, ozone (O_3) and SO_2 were introduced (see Extended Data Table 1). Using various instruments (Methods and Extended Data Fig. 1) we measured the behaviour of freshly nucleated particles of 1–2 nm diameter and their subsequent growth up to 80 nm. Two chemical ionization mass spectrometers (Methods) using nitrate as the reagent ion (nitrate-CI-API-TOF) measured the concentrations of sulfuric acid and highly oxygenated organic compounds^{16,17}. Nitrate anions tend to cluster with highly oxygenated molecules (HOMs), and the measured HOMs fall broadly into two product ranges based on carbon number (Extended Data Fig. 2): monomers ($C_xH_yO_z$ with $x = 8–10$, $y = 12–16$ and $z = 6–12$), and dimers ($C_xH_yO_z$ with $x = 17–20$, $y = 26–32$ and $z = 8–18$). Here we refer to these measured compounds as HOMs rather than extremely low-volatility organic compounds (ELVOCs), as previously reported¹⁷. As we shall show, the HOM volatility spans a wide range (although it is always very low), and we shall separate HOMs into volatility bins using the volatility basis set (VBS)¹⁸.

¹Paul Scherrer Institute, Laboratory of Atmospheric Chemistry, CH-5232 Villigen, Switzerland. ²Carnegie Mellon University, Center for Atmospheric Particle Studies, Pittsburgh, Pennsylvania 15213, USA. ³CERN, CH-1211 Geneva, Switzerland. ⁴Goethe University Frankfurt, Institute for Atmospheric and Environmental Sciences, 60438 Frankfurt am Main, Germany. ⁵Department of Physics, University of Helsinki, PO Box 64, FI-00014 Helsinki, Finland. ⁶Department of Applied Environmental Science, University of Stockholm, SE-10961 Stockholm, Sweden. ⁷Institute for Atmospheric and Climate Science, ETH Zürich, 8092 Zürich, Switzerland. ⁸Chemical Sciences Division, Earth System Research Laboratory, NOAA, Boulder, Colorado, USA. ⁹Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125, USA. ¹⁰Helsinki Institute of Physics, University of Helsinki, PO Box 64, FI-00014 Helsinki, Finland. ¹¹Institute for Ion and Applied Physics, University of Innsbruck, 6020 Innsbruck, Austria. ¹²Ionicon Analytik GmbH, 6020 Innsbruck, Austria. ¹³WSL Institute for Snow and Avalanche Research SLF, 7260 Davos, Switzerland. ¹⁴University of Eastern Finland, 70211 Kuopio, Finland. ¹⁵Finnish Meteorological Institute, 00101 Helsinki, Finland. ¹⁶National Center for Atmospheric Research, Atmospheric Chemistry Observations and Modeling Laboratory, Boulder, Colorado 80301, USA. ¹⁷Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology, Karlsruhe, Germany. ¹⁸School of Earth and Environment, University of Leeds, LS2 9JT Leeds, UK. ¹⁹Department of Chemistry, University of California, Irvine, California 92697, USA. ²⁰Faculty of Physics, University of Vienna, 1090 Vienna, Austria. ²¹SIM, University of Lisbon and University of Beira Interior, 1849-016 Lisbon, Portugal. ²²Aerodyne Research, Inc., Billerica, Massachusetts 01821, USA. [†]Present addresses: Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado, USA and Chemical Sciences Division NOAA Earth System Research Laboratory, Boulder, Colorado, USA (C.W.); SMEAR II, Hyttälä Forestry Field Station, University of Helsinki, Hyttäläntie 124, FI-35500 Korkeakoski, Finland (H.K.); Arctic Research Center, Korea Polar Research Institute, 21990 Incheon, South Korea (J. Kim); Department of Atmospheric Sciences, University of Washington, Seattle, Washington 98195, USA (S.S.); Institute for Aerosol and Sensor Technology, University of Applied Science Northwestern Switzerland, 5210 Windisch, Switzerland (E.W.).

In Fig. 1 we plot the growth rates measured in CLOUD as a function of sulfuric acid and HOM concentration, focusing on size ranges from 1.1 nm to 3.2 nm and >5 nm (mobility) diameter. It is evident from Fig. 1a and b that the observed growth cannot be explained in either size range by the condensation of sulfuric acid even at the kinetic limit, where sulfuric acid is assumed to be completely non-volatile. Furthermore, for sulfuric acid molecular concentrations below 10^7 cm^{-3} , the growth rate is uncorrelated with sulfuric acid. In contrast, the growth is clearly correlated with organics for all size ranges up to the size of cloud condensation nuclei (CCN) for HOM concentrations $>10^6 \text{ cm}^{-3}$ (Fig. 1c and d). For experiments with sulfuric acid concentration $<5.5 \times 10^5 \text{ cm}^{-3}$ we have separately reported a large charge enhancement for the nucleation rate¹⁵. However, there is no corresponding charge influence on the growth rates of either 1.1–3.2 nm or >5 nm particles (grey versus blue symbols in Fig. 1c and d). Most of the HOMs in the chamber are neutral ($\sim 10^7 \text{ cm}^{-3}$ neutral HOMs versus $\sim 10^3 \text{ cm}^{-3}$ charged molecules), so a charge enhancement is not expected, especially with increasing size¹⁹. However, owing to the experimental uncertainties we cannot exclude the possibility of an ion enhancement at sizes below 3 nm.

A non-volatile (collision-limited) model of HOM condensation (Methods) cannot explain the observed growth rates across the full range of particle diameters we studied. We modelled growth at 1.1 nm, 3.2 nm, 5 nm, 15 nm and 50 nm (labelled curves, Fig. 1c and d) assuming that observed HOM monomers and dimers are non-volatile, with a density of $1,400 \text{ kg m}^{-3}$ and a mass of 300 Da. Contrary to the common

misconception that non-volatile diameter growth rate should be constant with size (in the free molecular regime), the predicted growth rate with this assumption is highest at any given HOM concentration for the smallest particles and decreases rapidly with increasing size up to ~ 5 nm (Fig. 1c, d). This predicted decreasing growth rate with increasing particle size is because the cross-section and collision velocity are highest relative to particle size for the smallest particles (Methods). However, the observations show the reverse, with growth rates for sizes above 5 nm exceeding those near 2 nm by a factor of 1.5 ± 0.2 , obtained from normalizing (to 10^7 cm^{-3} HOMs) and averaging the growth rates in the considered size ranges. The ratio of observed growth rates to modelled non-volatile growth rates increases from 0.7 ± 0.1 at 1.1 nm to 2.8 ± 0.2 at 5 nm, where in each case the quoted error is the standard error of the mean. This large discrepancy is strong evidence that the measured HOMs cannot fully describe the observed growth, and that additional organic material must be contributing to particle growth above roughly 5 nm particle diameter.

To explore the potential role of HOM volatility, we use the SIMPOL model²⁰ to estimate the saturation mass concentration (C^* , $\mu\text{g m}^{-3}$) and saturation molecular concentration (N^* , cm^{-3}) of each HOM using its measured atomic composition together with an estimation of its likely chemical structure (see Extended Data Fig. 3). We grouped the HOMs in volatility bins (separated by factors of ten) and assigned them to several volatility classes (see Extended Data Fig. 4). The HOMs span a wide range from extremely low-volatility (ELVOC, $C^* < 10^{-4.5} \mu\text{g m}^{-3}$; $N^* < 5 \times 10^4 \text{ cm}^{-3}$ assuming a molecular mass

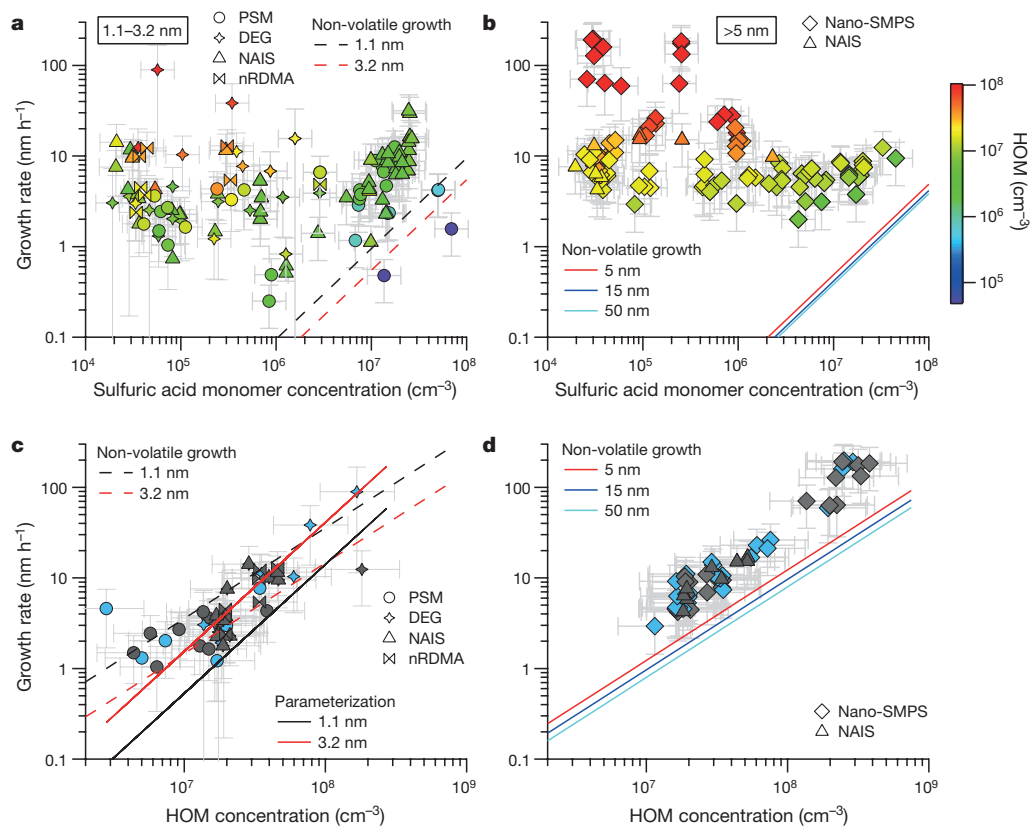


Figure 1 | Growth rates as a function of sulfuric acid and highly oxygenated molecule (HOM) concentrations. Symbol shapes represent the different instruments to derive the growth rates (see key and Methods), symbol colours indicate the HOM concentration (colour scale at right). **a, b**, Growth rate versus sulfuric acid concentration for particles from 1.1 nm to 3.2 nm (**a**), and for particles 5–15 nm, 15–30 nm, 30–60 nm and >60 nm (**b**). Non-volatile growth rates by condensation of sulfuric acid⁵ are displayed for different diameters. **c**, Measured growth rates from 1.1 nm to 3.2 nm versus the HOM concentration for sulfuric acid concentrations $<5 \times 10^5 \text{ cm}^{-3}$; **d**, as **c** but for size ranges 5–15 nm,

15–30 nm, 30–60 nm and >60 nm. Linear growth was observed for particles >5 nm, thus no differentiation was made in **b** and **d**. Panel **c** additionally shows the parameterization for 1.1 nm and 3.2 nm based on our volatility-distribution modelling results. Symbol colours in **c** and **d** indicate the ion conditions in the chamber (blue, neutral; grey, ions from Galactic cosmic rays (GCR); see Methods). The HOM and sulfuric acid concentration uncertainty (error bars) is estimated to be $+80\%/-45\%$ and $+50\%/-33\%$, respectively. Growth rate error bars indicate 1σ total errors.

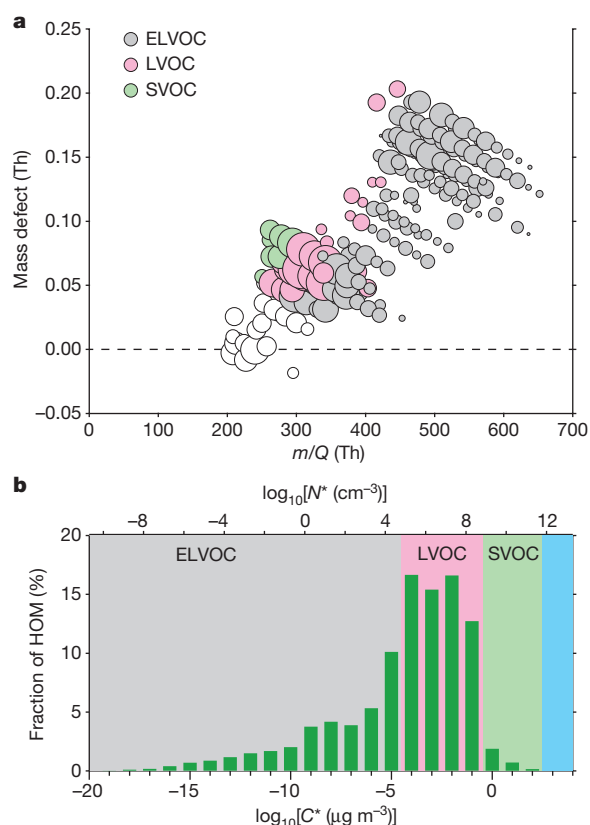


Figure 2 | Observed gas-phase HOMs and their volatility distribution. **a**, Mass defect (in Th; 1 Th = 1 Da/ e , where e is the elementary charge) of all HOMs versus their nominal mass to charge ratio (m/Q) including the estimated volatility distribution based on the proposed structures (Extended Data Fig. 3). The size of the plotting symbols is proportional to the logarithm of the counting rate. White circles are C_5 – C_7 compounds, which were not included in the volatility analysis. **b**, HOMs binned to a volatility distribution showing the measured relative counting rates in per cent, with ELVOCs comprising ~36%.

of 300 Da) to low-volatility (LVOC, $10^{-4.5} \mu\text{g m}^{-3} \leq C^* \leq 10^{-0.5} \mu\text{g m}^{-3}$; $5 \times 10^4 \text{ cm}^{-3} \leq N^* \leq 5 \times 10^8 \text{ cm}^{-3}$) to some semi-volatile (SVOC, $10^{-0.5} \mu\text{g m}^{-3} \leq C^* \leq 10^{2.5} \mu\text{g m}^{-3}$; $5 \times 10^8 \text{ cm}^{-3} \leq N^* \leq 5 \times 10^{11} \text{ cm}^{-3}$) organic compounds. In Fig. 2a we show a mass defect plot (Methods) of the observed compounds during a representative run, and in Fig. 2b we show the corresponding volatility distribution (colours based on ref. 18). The binned volatility distribution of measured gas-phase organic species (Fig. 2b) shows a substantial fraction of ELVOCs, maximal contribution in the LVOC range and even low levels of SVOCs. Because the LVOCs and SVOCs do not build up a sufficient saturation ratio to overcome the Kelvin barrier, they should not be able to condense onto the smallest particles, so that only the ELVOCs should contribute to the initial growth. While nitrate ions cluster efficiently with ELVOCs and calibration based on sulfuric acid should be fairly accurate, the concentration of LVOCs and SVOCs is likely to be underestimated because of inefficient clustering²¹. Indeed, SVOCs are formed with high yield in α -pinene oxidation²² but most of them evidently are not detected by the nitrate-CI-API-TOF instrument (Fig. 2). The fact that even the non-volatile model based on measured HOMs underestimates the observed growth rates for particles >5 nm by a factor of three strongly indicates that the concentration of condensing organic vapours is substantially higher than measured, at least after the Kelvin barrier has diminished.

We further consider two very different experiments. During the first experiment, the HOM concentration increased nonlinearly with time, which replicates the diurnal variation of biogenic emissions and oxidants in the ambient for the morning and early afternoon (Fig. 3a). This situation leads to a nonlinear increase in the growth rate. During the second experiment, the HOM concentration remained at a constant steady state (production balanced by wall loss). This allowed us to test whether the accelerating growth seen in the first experiment was due to the diminishing Kelvin effect or the increasing HOM concentration. The constant HOM concentration led to a nearly constant growth rate, except for the smallest particles below ~ 5 nm (Fig. 3d).

In order to quantify the importance of the Kelvin effect and HOM measurement biases, we analysed the contribution of HOMs to early growth and assessed the dependence on HOM volatility by using a dynamic volatility-distribution model²³ for these two cases. The HOM volatility-distribution model comprises nine C^* bins ranging from

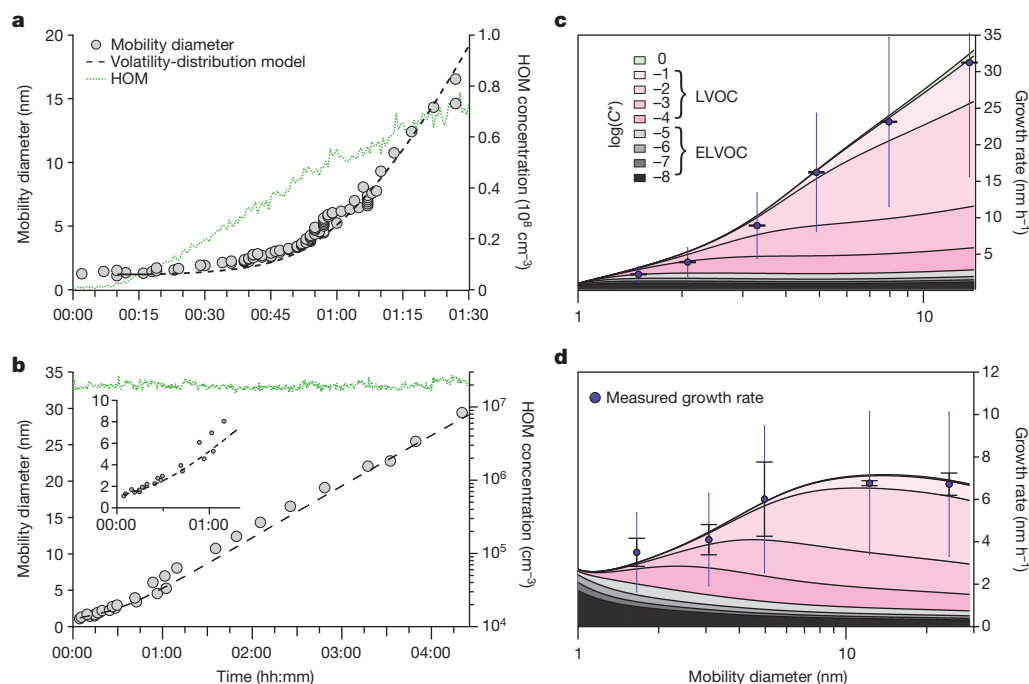


Figure 3 | Comparison of the growth rates in two experiments with a dynamic volatility basis set (VBS) model. **a**, **b**, Temporal evolution of the particle size (filled circles) and the modelled particle size (dashed lines) for an experiment with increasing HOM concentration (**a**), and for constant HOM concentration (**b**), with the inset magnifying the time evolution of the first 5 nm. **c**, **d**, Size-dependent modelled (lines) and measured (filled circles) growth rate for the increasing HOM concentration (**c**), and for the constant HOM concentration (**d**). Colours (key in **c**) indicate the contribution of different volatility bins to the condensational growth. Error bars indicate the error of the fit alone, whiskers the 1σ systematic scale uncertainty of the determined growth rates.

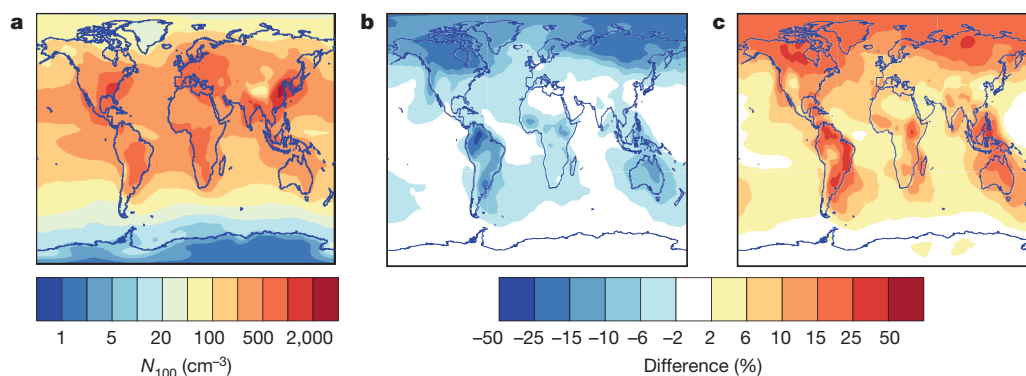


Figure 4 | Modelled influence on global CCN of different organic growth rates from 1.7 nm to 3 nm simulated by the GLOMAP aerosol model. a, The annual mean number concentration of soluble particles of at least 100 nm diameter (N_{100} on colour scale, taken as a proxy for CCN) at cloud base level. We treat irreversible (collision-limited) condensation of sulfuric acid for particle growth from 1.7 nm to 3 nm, together with a size-dependent growth rate due to HOMs from the present work.

$10^{-8} \mu\text{g m}^{-3}$ to $1 \mu\text{g m}^{-3}$ (10^1 cm^{-3} to 10^9 cm^{-3}), split into three ranges (see Fig. 2 and Extended Data Fig. 5): ELVOC (grey), LVOC (pink) and SVOC (light green). When we run the HOM volatility-distribution model using the directly measured volatility-binned HOM concentrations as input, the simulated growth rates for particles >2 nm are underestimated by a large factor (see Extended Data Fig. 6, blue dashed line). This is consistent with the expectation that the detection efficiency of LVOCs in the nitrate-CI-API-TOF is lower as discussed above. An attempt to adjust the HOM volatility distribution by increasing the LVOCs to reproduce the observed growth rates was not successful (see Extended Data Fig. 6, blue solid line). The model can be brought into agreement with observations by increasing the LVOC concentrations and introducing a Kelvin effect (Fig. 3 and Extended Data Fig. 6 grey line). This tuned model, adjusting for inefficient LVOC measurement in the nitrate-CI-API-TOF and considering the Kelvin effect (see Methods, Extended Data Fig. 5b and Extended Data Fig. 7 for details), captures the observed particle growth in both example cases with high fidelity (Fig. 3). While the agreement at 10 nm diameter is ensured by our LVOC correction, the Kelvin term is essential to reproduce the observed growth rate over the full size range for these two quite different cases, although the strong size dependence in Fig. 3a is primarily due to the increasing HOM concentration. This is evidence that the Kelvin term (along with abundant LVOCs) is responsible for the acceleration in growth observed in field experiments in the afternoon, and that only ELVOCs have a sufficiently high saturation ratio to overcome the Kelvin barrier at the smallest sizes.

The pool of ELVOCs, many having $C^* \ll 10^{-8} \mu\text{g m}^{-3}$ (Fig. 2b), implies continuous production of relatively stable clusters smaller than 2 nm (continuous nucleation is observed, as shown in Extended Data Fig. 8). ELVOCs govern the contribution to growth up to ~ 2 nm; beyond this, LVOCs take over in sequence as the Kelvin effect becomes progressively weaker with increasing size. Thus, while growth rates in the non-volatile HOM model decrease by a factor of ~ 3 between 1 nm and 5 nm, in the volatility-distribution HOM model they increase by a factor of ~ 3 over this range, consistent with observations. This volatility-distribution growth model is a version of ‘nano-Köhler theory’, in which the effects of condensed-phase mixing (Raoult’s law) and particle curvature (the Kelvin term) combine for miscible organics. The Kelvin effect dominates because curvature enhances condensed-phase activities by orders of magnitude for the smallest particles, regardless of their composition, and the critical issue is whether the saturation ratio of an LVOC volatility bin exceeds this threshold (see Extended Data Fig. 7 for detailed model results). Finally, the volatility-distribution model shows that, in the experiments, SVOCs cannot contribute to the observed growth via non-reactive uptake as their gas-phase saturation ratio never

b, The percentage change in CCN concentration (colour scale) when growth from 1.7 nm to 3 nm is due to sulfuric acid alone. **c,** The percentage change in CCN concentration when we parameterize growth from 1.7 nm to 3 nm as irreversible condensation of sulfuric acid together with an organic contribution following ref. 30, which assumes a Kelvin barrier to organic condensation below 2.5 nm. All simulations assume the same nucleation rates at 1.7 nm and the same particle growth rates above 3 nm.

rises high enough for them to contribute (see Extended Data Fig. 7 and Methods).

The α -pinene + ozone system explored here is among the most efficient sources of ELVOCs yet observed^{16,17}, but it is likely that many sources of LVOCs may be important in the atmosphere. The latter include the first-generation compounds described here but also later-generation ‘ageing’ products formed by reaction with OH radicals^{10,24,25}. Different sources are almost certain to produce LVOCs with differing volatility distributions and chemical properties, which also might influence their reactivity in the condensed phase, including oligomerization²³ and reactive uptake²⁶, resulting in different growth patterns compared to those in Fig. 3. These growth patterns thus constitute a critical and variable link between new particle formation and CCN formation.

Strongly size-dependent nanoparticle growth has been observed and parameterized based on atmospheric observations^{3,27–29}, although during nucleation events in the field it has not been possible to determine whether changes in the growth rate are due to the Kelvin effect or due to changes in the HOM concentrations during the event. To assess the global implications of our findings, we parameterized the growth between 1.7 nm and 3 nm using the size-resolved growth rates from the HOM volatility-distribution modelling results (Fig. 1 and Methods). Using a global aerosol model (Methods), we find that CCN concentrations are sensitive to whether, and how, organic compounds participate in the first stages of the growth of freshly nucleated particles. Figure 4a shows the concentrations of soluble 100 nm particles (N_{100}), a proxy for CCN, using our parameterized growth rates, which are up to a factor of two higher than those in a simulation without organics participating in the initial growth (Fig. 4b). Conversely, a previous parameterization³⁰ which empirically accounts for the Kelvin effect below 2.5 nm but assumes that all condensable organic products (not just HOMs) contribute to the growth of these particles, produces CCN concentrations up to 50% higher than our parameterization (Fig. 4c). Our model results show that CCN concentrations can be sensitive to the processes and concentrations of species driving the growth of the smallest atmospheric particles as reflected in the pronounced differences of the corresponding growth rates (Extended Data Fig. 9). On the basis of the combined modelling results and experimental data that we report here, we suggest that low-volatility organic vapours are the key to particle growth at the initial sizes.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.


Received 1 September 2015; accepted 22 April 2016.

1. Merikanto, J., Spracklen, D. V., Mann, G. W., Pickering, S. J. & Carslaw, K. S. Impact of nucleation on global CCN. *Atmos. Chem. Phys.* **9**, 8601–8616 (2009).
2. Kulmala, M. *et al.* Direct observations of atmospheric aerosol nucleation. *Science* **339**, 943–946 (2013).
3. Kuang, C. *et al.* Size and time-resolved growth rate measurements of 1 to 5 nm freshly formed atmospheric nuclei. *Atmos. Chem. Phys.* **12**, 3573–3589 (2012).
4. Lehtinen, K. E., Dal Maso, M., Kulmala, M. & Kerminen, V.-M. Estimating nucleation rates from apparent particle formation rates and vice versa: revised formulation of the Kerminen-Kulmala equation. *J. Aerosol Sci.* **38**, 988–994 (2007).
5. Nieminen, T., Lehtinen, K. E. J. & Kulmala, M. Sub-10 nm particle growth by vapor condensation-effects of vapor molecule size and particle thermal speed. *Atmos. Chem. Phys.* **10**, 9773–9779 (2010).
6. Riccobono, F. *et al.* Contribution of sulfuric acid and oxidized organic compounds to particle formation and growth. *Atmos. Chem. Phys.* **12**, 9427–9439 (2012).
7. Riipinen, I. *et al.* The contribution of organics to atmospheric nanoparticle growth. *Nat. Geosci.* **5**, 453–458 (2012).
8. Smith, J. N. *et al.* Chemical composition of atmospheric nanoparticles formed from nucleation in Tecamac, Mexico: evidence for an important role for organic species in nanoparticle growth. *Geophys. Res. Lett.* **35**, L04808 (2008).
9. Laaksonen, A. *et al.* The role of VOC oxidation products in continental new particle formation. *Atmos. Chem. Phys.* **8**, 2657–2665 (2008).
10. Donahue, N. M. *et al.* How do organic vapors contribute to new-particle formation? *Faraday Discuss.* **165**, 91–104 (2013).
11. Zhao, J., Ortega, J., Chen, M., McMurtry, P. & Smith, J. Dependence of particle nucleation and growth on high-molecular-weight gas-phase products during ozonolysis of α -pinene. *Atmos. Chem. Phys.* **13**, 7631–7644 (2013).
12. Donahue, N. M., Trump, E. R., Pierce, J. R. & Riipinen, I. Theoretical constraints on pure vapor-pressure driven condensation of organics to ultrafine particles. *Geophys. Res. Lett.* **38**, L16801 (2011).
13. Pierce, J. R. *et al.* Quantification of the volatility of secondary organic compounds in ultrafine particles during nucleation events. *Atmos. Chem. Phys.* **11**, 9019–9036 (2011).
14. Kulmala, M., Kerminen, V.-M., Anttila, T., Laaksonen, A. & O'Dowd, C. D. Organic aerosol formation via sulphate cluster activation. *J. Geophys. Res.* **D 109**, D04205 (2004).
15. Kirkby, J. *et al.* Ion-induced nucleation of pure biogenic particles. *Nature* **533**, <http://dx.doi.org/10.1038/nature17953> (2016).
16. Jokinen, T. *et al.* Production of extremely low volatile organic compounds from biogenic emissions: measured yields and atmospheric implications. *Proc. Natl Acad. Sci. USA* **112**, 7123–7128 (2015).
17. Ehn, M. *et al.* A large source of low-volatility secondary organic aerosol. *Nature* **506**, 476–479 (2014).
18. Donahue, N. M., Kroll, J. H., Pandis, S. N. & Robinson, A. L. A two-dimensional volatility basis set — part 2: diagnostics of organic-aerosol evolution. *Atmos. Chem. Phys.* **12**, 615–634 (2012).
19. Lovejoy, E. R., Curtius, J. & Froyd, K. D. Atmospheric ion-induced nucleation of sulfuric acid and water. *J. Geophys. Res.* **D 109**, D08204 (2004).
20. Pankow, J. F. & Asher, W. E. SIMPOL. 1: A simple group contribution method for predicting vapor pressures and enthalpies of vaporization of multifunctional organic compounds. *Atmos. Chem. Phys.* **8**, 2773–2796 (2008).
21. Hyttinen, N. *et al.* Modeling the charging of highly oxidized cyclohexene ozonolysis products using nitrate-based chemical ionization. *J. Phys. Chem. A* **119**, 6339–6345 (2015).
22. Presto, A. A. & Donahue, N. M. Investigation of α -pinene + ozone secondary organic aerosol formation at low total aerosol mass. *Environ. Sci. Technol.* **40**, 3536–3543 (2006).
23. Trump, E. R. & Donahue, N. M. Oligomer formation within secondary organic aerosols: equilibrium and dynamic considerations. *Atmos. Chem. Phys.* **14**, 3691–3701 (2014).
24. Schobesberger, S. *et al.* Molecular understanding of atmospheric particle formation from sulfuric acid and large oxidized organic molecules. *Proc. Natl Acad. Sci. USA* **110**, 17223–17228 (2013).
25. Riccobono, F. *et al.* Oxidation products of biogenic emissions contribute to nucleation of atmospheric particles. *Science* **344**, 717–721 (2014).
26. Wang, L. *et al.* Atmospheric nanoparticles formed from heterogeneous reactions of organics. *Nat. Geosci.* **3**, 238–242 (2010).
27. Yli-Juuti, T. *et al.* Growth rates of nucleation mode particles in Hyytiälä during 2003–2009: variation with particle size, season, data analysis method and ambient conditions. *Atmos. Chem. Phys.* **11**, 12865–12886 (2011).
28. Häkkinen, S. A. K. *et al.* Semi-empirical parameterization of size-dependent atmospheric nanoparticle growth in continental environments. *Atmos. Chem. Phys.* **13**, 7665–7682 (2013).
29. Bianchi, F. *et al.* New particle formation in the free troposphere: a question of chemistry and timing. *Science* **352**, <http://dx.doi.org/10.1126/science.aad5456> (2016).
30. D'Andrea, S. D. *et al.* Understanding global secondary organic aerosol amount and size-resolved condensational behavior. *Atmos. Chem. Phys.* **13**, 11519–11534 (2013).

Acknowledgements We thank CERN for supporting CLOUD with technical and financial resources, and for providing a particle beam from the CERN Proton Synchrotron. This research has received funding from the EC Seventh Framework Programme (Marie Curie Initial Training Network 'CLOUD-ITN' no. 215072, MC-ITN 'CLOUD-TRAIN', no. 316662, and ERC-StG-ATMOGAIN (278277) and ERC-Advanced 'ATMNUCLE' grant no. 227463), the German Federal Ministry of Education and Research (project nos 01LK0902A and 01LK1222A), the Swiss National Science Foundation (project nos 200020_135307, 200020_152907, 20FI20_149002 and 200021_140663), the Academy of Finland Center of Excellence programme (project no. 1118615), the Academy of Finland (CoE project no. 1118615, LASTU project no. 135054), the Næssing Foundation, the Austrian Science Fund (FWF; project no. J3198-N21), the EU's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie (no. 656994), the Swedish Research Council, Vetenskapsrådet (grant no. 2011-5120), the Portuguese Foundation for Science and Technology (project no. CERN/FP/116387/2010), the Presidium of the Russian Academy of Sciences and Russian Foundation for Basic Research (grants 08-02-91006-CERN and 12-02-91522-CERN), Dreyfus Award EP-11-117, the Davidow Foundation, the US National Science Foundation (grants AGS1136479, AGS1447056, AGS1439551 and CHE1012293), US Department of Energy (grant DE-SC00014469) and the FP7 project BACCHUS (grant agreement 603445).

Author Contributions A.A., J.A., U.B., A.-K.B., M.B., F.B., K.S.C., J.S.C., J.C., A.D., J.Do., N.M.D., J.Du., S.E., R.C.F., A.F., C.Fr., C.Fu., R.G., M.G., M.H., T.J., H.K., J.Kir., M.Kr., M.Ku., A.K., A.L., K.L., P.M., U.M., T.N., T.P., F.M.P., M.P.R., S.S., M.Sim., M.Sip., J.N.S., G.S., A.T., J.T., A.V., A.C.W., R.W., E.W., D.W., P.M.W., D.W. and C.Y. designed the experiment or prepared the CLOUD facility or instruments. A.A., J.A., A.-K.B., M.B., F.B., S.B., J.S.C., J.C., A.D., J.Du., S.E., A.F., C.Fr., C.Fu., H.G., M.H., C.R.H., T.J., J.Ka., H.K., J. Kim, J.Kir., M.Kr., A.K., M.L., K.L., P.M., U.M., T.N., F.M.P., I.R., M.P.R., N.S., S.S., K.S., M.Sim., M.Sip., J.N.S., G.S., A.T., J.T., A.V., A.C.W., R.W., C.W., D.W., C.Y. and P.Y. collected data. L.A., A.K.B., F.B., S.B., J.S.C., N.M.D., R.C.F., A.F., C.F., M.H., C.R.H., T.J., K.L., U.M., T.N., N.S., S.S., M.Sim., M.Sip., G.S., J.T., R.W., C.W., D.W. and C.Y. performed data analysis. J.Do. and U.M. contributed HOM structures. W.K.C., N.M.D., L.A., I.R. and J.T. performed aerosol growth modelling. H.G. performed GLOMAP modelling. J.T., L.A., U.B., F.B., K.S.C., J.C., J.Do., N.M.D., J.Du., R.C.F., C.Fr., H.G., M.G., M.H., C.R.H., T.J., J.Kir., M.Ku., K.L., U.M., T.P., I.R., M.P.R., N.S., S.S., M.Sim., C.W., D.W. and C.Y. were involved in the scientific interpretation and discussion. J.T., U.B., J.Do., N.M.D. and H.G. wrote the manuscript. All commented on the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the article. Correspondence and requests for materials should be addressed to U.B. (urs.baltensperger@psi.ch).

 This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

METHODS

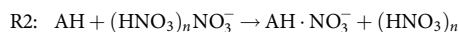
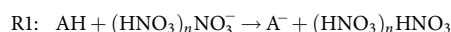
The CLOUD chamber. We conducted two measurement campaigns at the CERN CLOUD chamber, a 26 m³ stainless steel vessel which enables aerosol experiments under the full range of tropospheric conditions^{31,32}. CLOUD7, in the autumn of 2012, included mostly high sulfuric acid concentrations, while CLOUD8, in 2013, addressed low sulfuric acid concentrations. To avoid contamination, pure air is generated by the evaporation of cryogenic liquid nitrogen (N₂) and liquid oxygen (O₂), combined at a ratio of 79:21. A UV light (250–400 nm) system enables the formation of hydroxyl (OH) radicals via the photolysis of ozone³³. By applying a high voltage field (30 kV m⁻¹) all charged particles in the chamber can be removed rapidly (neutral conditions); when the high voltage field is turned off, natural ions are produced in the chamber by Galactic cosmic rays (GCR condition) reaching ground level. With the 3.5 GeV/c secondary pion beam (π condition) from the CERN Proton Synchrotron passing through the chamber, ion concentrations representative for those of the upper troposphere can be achieved^{31,34}. A dedicated inlet system is available for every gas. In order to clean the chamber, the chamber can be heated by raising the temperature to 373 K, and, in addition, flushed with ultra pure water. All gas pipes are made from stainless steel to avoid contamination, and chamber and gas seals are chemically inert gold coated metal. Two fans running in counter flow ensure a good mixture of the gases in the chamber³⁵. Traces of contaminants, for example, low molecular weight volatile organic compounds (VOCs)³⁶ or ammonia³⁷, were sometimes observed in the chamber. However, as shown elsewhere³⁶, extremely clean conditions can be achieved.

Experimental settings. A typical experiment started with the injection of α -pinene under neutral (ion free) conditions. The ozone already present in the chamber immediately reacts with the α -pinene leading to aerosol nucleation (see also ref. 15). Using the UV fibre system in the chamber, additional OH could be photochemically produced. The major fraction of HOM (~60%) were chemically produced via the ozonolysis of α -pinene. This experiment was continued until a steady-state—that is, a stable HOM concentration, was achieved. Afterwards the high-voltage field, used in neutral experiments for sweeping out ions, was turned off. This allowed ions (~700 cm⁻³) produced by Galactic cosmic rays to accumulate in the chamber, and resulted in a second nucleation event (see also ref. 15). In addition experiments were also started under GCR conditions to prove consistency. In total, approximately 40% of the runs started (with increasing HOM concentration) in neutral conditions, 18% in GCR condition and 20% in π condition. Plateau conditions (with steady-state HOM concentration) in GCR constitute approximately 18% of the runs and in π condition approximately 4%. π conditions relate to experiments where the Proton Synchrotron was also used to produce higher ion concentrations (~3,000 cm⁻³), as encountered in the upper troposphere. This was only possible during CLOUD 7, as during CLOUD 8 the Proton Synchrotron was not in operation due to maintenance work. A typical experiment is shown in Extended Data Fig. 8. For pure biogenic experiments, we added no SO₂; for sulfuric acid experiments, we injected SO₂ into the chamber as an additional precursor. All experimental steady-state conditions can be found in Extended Data Table 1. For each run several growth rates at different diameters could be quantified (see Extended Data Figs 1 and 8). Extended Data Fig. 8 shows two nucleation events that were observed during one run, one under neutral and the second one under GCR conditions. Thus, one run can yield several points in Fig. 1.

Cluster composition. Atmospheric pressure interface time of flight mass spectrometer (API-TOF). One API-TOF (ToFwerk AG) measured the mass-to-charge ratio of positive or negative clusters present in the CLOUD chamber²⁴. Since this instrument only measures charged clusters, the measurements were made during GCR or π conditions. It is only possible to measure one polarity at a time thus positive and negative spectra were measured alternately.

Chemical ionization atmospheric pressure interface time of flight mass spectrometer (nitrate-CI-API-TOF). Two nitrate-CI-API-TOFs³⁸ measured the concentration of sulfuric acid, oxidized organics and other clusters and molecules in the cloud chamber.

The instruments use an ion source (one an X-ray generator, one a corona needle) to ionize the reagent gas nitric acid in a nitrogen flow. In a drift tube an electric field is then applied to guide the primary ions to the sample flow where they react with the neutral molecules and clusters with an overall reaction time of about 200 ms in one instrument and 50 ms in the other. Inside the drift tube, two possible reactions can then take place to ionize the neutral molecules or clusters A in the sample flow:



The first reaction (R1) corresponds to a proton transfer reaction (acid/base reaction) which is, for example, the case for sulfuric acid. The second reaction (R2) is a ligand switching reaction, forming a more stable adduct, which is the case for

highly oxygenated molecules (HOMs). Using an electrostatic field, the charged molecules and clusters are then guided through a small pinhole with a diameter of 350 (300) μm to the API-TOF section.

The voltage settings in the API and TOF sections determine the mass dependent transmission efficiency of the instrument. The transmission curves were determined with separate measurements, by adding certain compounds (perfluorinated acids) to the instrument in sufficient amounts to deplete the primary ions. With this method the transmission relative to the mass to charge ratio (m/Q) of the primary ions was determined³⁹. One instrument operated at the same voltage settings for the whole campaign while the other one was operated in a switching mode between voltage settings optimized for a low or high m/Q range.

We analysed the raw data with the MATLAB *tofTools* package^{32,40}. The mass scale is calibrated to better than 10 p.p.m. accuracy, using a two-parameter fit. The concentration of sulfuric acid is calculated from the count rates of each ion species as follows:

$$[\text{H}_2\text{SO}_4] = C \times \text{SL}_{\text{H}_2\text{SO}_4} \times \ln \left(1 + \frac{[\text{HSO}_4^-] + [\text{HSO}_4^- \cdot \text{HNO}_3]}{\sum_{j=0}^2 [\text{NO}_3^- \cdot (\text{HNO}_3)_j]} \right) \quad (1)$$

where $[\text{H}_2\text{SO}_4]$ is the concentration of sulfuric acid. The corresponding ion count rates, including the primary ions, appear on the right hand side of the equation. C is a calibration coefficient, which was determined by connecting the nitrate-CI-API-TOF to a well characterized H₂SO₄ generator⁴¹. Line losses for H₂SO₄ were corrected with the term $\text{SL}_{\text{H}_2\text{SO}_4}$. SL can be calculated from empirical equations for straight circular tubes with a laminar flow⁴².

Measurement of oxidized organics. During nucleation and growth, we observed two distinct signal patterns—monomers and dimers—in the nitrate-CI-API-TOF (Extended Data Fig. 2, Run 1209) corresponding to the monomers and dimers of the α -pinene oxidation products. These bands contain highly oxygenated molecules (HOMs), which have been found to play a potentially major role in aerosol growth¹⁷. Owing to their structure and their high O:C, these clusters have a low saturation vapour concentration. In ref. 17, it was assumed that all observed oxygenated organics are extremely low-volatility organic compounds (ELVOCs) and condense on the added seed aerosol.

We define monomers (mainly C_xH_yO_z with $x = 8-10$, $y = 12-16$ and $z = 6-12$) as the sum of the peaks in the m/Q range from 235–424 Th (1 Th = 1 Da/ e , where e is the elementary charge) and dimers (mainly C_xH_yO_z with $x = 17-20$, $y = 26-32$ and $z = 8-18$) as the sum from 425–625 Th. We excluded contamination peaks from the summation within a band, as well as peaks assigned to RO₂ radicals (C₁₀H₁₅O_{6,8,10,12}, corresponding to m/Q of 293, 325, 357 and 389 Th).

The API-TOF also detected naturally charged clusters between 670 and 850 Th (trimers) and between 900 and 1,200 Th (tetramers). For the nitrate-CI-API-TOF the trimer band was only observed for a very long integration time, indicating either a low concentration of neutral trimers or a low transmission efficiency. We also observed intermediate species with a carbon number of 11 to 17, which may be dimers formed from reactions of RO₂ radicals with RO₂ radicals formed from fragments. However, their concentration is small (see cyan peaks in Extended Data Fig. 2).

To estimate the concentration of each highly oxygenated molecule (HOM_{*i*}), we applied the following equation:

$$[\text{HOM}_i] = C \times T_i \times \text{SL}_{\text{HOM}_i} \times \ln \left(1 + \frac{[\text{HOM}_i \cdot \text{NO}_3^-]}{\sum_{j=0}^2 [\text{NO}_3^- \cdot (\text{HNO}_3)_j]} \right) \quad (2)$$

In this equation, $[\text{HOM}_i \cdot \text{NO}_3^-]$ is the integrated area of a background corrected HOM peak in counts per second (c.p.s.). We corrected for the losses through the sampling line with the term SL_{HOM} . Here, we used the diffusion coefficients for the monomers (0.0297 cm² s⁻¹) and for the dimers (0.0240 cm² s⁻¹), which we determined in the CLOUD chamber experimentally. This results in correction factors for the monomers of a factor of 1.44 and for dimers of a factor of 1.37. The total HOM concentration is defined as the sum of all $[\text{HOM}_i]$, which includes all identified monomers, dimers and intermediate clusters (see Extended Data Fig. 2).

We assume that the binding between the nitrate ion and the HOM is strong and proceeds at the kinetic limit and therefore use the same calibration constant C as for sulfuric acid. This assumption does hold for highly oxygenated species with extremely low volatilities, but not for less oxygenated species as the ionization efficiency decreases²¹. Quantum chemical calculations have shown that the nitrate

preferably clusters with ELVOC²¹. Less oxidized species are, therefore, observed to a lesser extent under our experimental conditions (HNO₃ concentration).

The transmission efficiency T_i of each individual HOM_{*i*} depends strongly on the mass of each molecule and the different voltage settings in the nitrate-CI-API-TOF. To correct this transmission factor, we derived a transmission curve over the whole mass range of the HOMs. For more details see ref. 43.

The uncertainty in HOM measurement was caused by the following sources: uncertainty in sulfuric acid calibration, charging efficiency of HOMs by the nitrate ion, mass dependent transmission efficiency and sampling line losses. This results in an overall scaling uncertainty for the measured [ELVOC] of +80%/–45% assuming a charging efficiency of one. We cannot give an uncertainty of the LVOC concentration. Instead we used a scaling factor to match the observation. On the basis of that and because LVOC \gg ELVOC, the HOM concentration is presumably underestimated by a factor of four. Nobody, at least to our knowledge, has been able to calibrate the nitrate chemical ionization source for charging efficiency so far.

For the analysis, the data from only one nitrate-CI-API-TOF (University of Frankfurt–UFRA) was used. The main reason for this was that a transmission calibration of the API-TOF section was performed with this instrument (see also ref. 43) and thus the data are expected to be quantitatively correct. The other nitrate-CI-API-TOF (University of Helsinki–UHEL) agrees very well for the monomer concentration, but less well for the oligomers. In addition, the UHEL nitrate-CI-API-TOF was operated under different settings. It was switched between several modes—(1) high fragmentation, (2) high mass and (3) low mass—to get further information on the fragmentation of the molecules and clusters.

Mass defect. In a mass defect plot, the difference between the exact mass of a compound and its nominal mass (M_n) is depicted as function of its mass to charge ratio (m/z). Depending on the element the mass defect can be negative or positive. In case of oxygen the mass defect is negative, so that a slope downwards represents an increase in oxygen molecules. Thus, the analysis of a complex high resolution spectrum is simplified by a convenient visualization where the pattern of compounds belonging to the same family is clearly shown.

Proton transfer reaction time of flight mass spectrometer (PTR-TOF-MS). We used a PTR-TOF-MS (Ionicon Analytik) to determine α -pinene concentrations in the chamber; it also provides information about the overall cleanliness regarding VOCs in the chamber. VOCs are ionized in a reaction chamber by means of a proton transfer reaction under precisely defined conditions (reaction time, pressure, temperature) and then analysed by a time-of-flight (TOF) mass spectrometer (Tofwerk AG). A mass resolving power of 4,000 ($m/\Delta m$, FWHM) and a mass accuracy within 10 p.p.m. enables unambiguous identification of pure hydrocarbons and volatile organic compounds up to $m/Q = 250$ Th (ref. 39). Direct calibration allows determination of α -pinene volume mixing ratios with an accuracy of 5% and a lower detection limit of 25 parts per trillion by volume (p.p.t.v.).

SO₂ chemical ionization mass spectrometer (SO₂-CIMS). The very low SO₂ volume mixing ratios were determined with an SO₂ chemical ionization mass spectrometer (SO₂-CIMS). It uses the primary ion (CO₃⁺) to convert SO₂ to SO₃⁺ (reaction scheme can be found elsewhere⁴⁴). The SO₃⁺ is then measured in a quadrupole mass spectrometer with an atmospheric pressure interface (Georgia Tech). The primary ions are generated with a corona discharge⁴⁵. The ratio of CO₃⁺ to NO₃⁺ was maximized by feeding CO₂, O₂ and Ar directly over the corona discharge, leading to a reduced contamination by NO₃⁺. The SO₂ concentration is then calculated as follows:

$$[\text{SO}_2] = C_s \ln(1 + R_{112}/R_{60}) \quad (3)$$

where R_{112} is the background-corrected ion count rate of SO₃⁺, R_{60} the ion count rate of CO₃⁺ and C_s the calibration factor. C_s was obtained by using an SO₂ gas standard (Carbagas AG). The calibration was repeated periodically during the campaign. The resulting calibration factor was found to be 1.3×10^5 p.p.t.v. Its detection threshold of SO₂ is about 15 p.p.t.v.; the uncertainty is within 23% for low SO₂ volume mixing ratios (around 30 p.p.t.v.), and 13% for volume mixing ratios >150 p.p.t.v. This uncertainty is mostly related to temperature changes in the experimental hall where the SO₂-CIMS was located. This change led to a drift in the SO₃⁺ background signal.

Aerosol properties. *Nano radial differential mobility analyser (nRDMA).* A custom-built aerosol size classifier and counter assembly was used to measure positively charged particles in the 1.1 to 10 nm diameter size range with a time resolution of 60 s. The classifier was a Caltech Nano-Radial Differential Mobility Analyser (herein referred to as nRDMA⁴⁶). The counter that was employed downstream of the nRDMA was an Airmodus Particle Size Magnifier with a 78 °C saturator coupled to a Brechtel Manufacturing Inc. Mixing Condensation Particle Counter, model 1710⁴⁷. The raw data from the Caltech assembly was inverted using transfer function parameters, effective length, and penetration efficiency functions⁴⁸.

Nano scanning mobility particle sizer (nano-SMPS). The nano-SMPS⁴⁹ measured the dry aerosol size distribution from 5 nm to 80 nm with a time resolution of 130 s. It was located within a temperature controlled rack and was kept at chamber temperature. The nano-SMPS consisted of the TSI condensation particle counter (CPC) 3772 with a modified cut-off ($D_{50} = 5.6$ nm, $D_{10} = 3.5$ nm)⁶, a TSI-type PSI-built short differential mobility analyser (DMA) and a neutralizer (Kr-85 source). The data were corrected for single charging efficiency, multiple charges, diffusion losses, and CPC detection efficiency. The diffusion loss correction assumes a laminar flow⁵⁰ and includes all parts of the nano-SMPS system (tubes, Kr-source, DMA inlet, DMA column).

Neutral cluster and air ion spectrometer (NAIS). The NAIS (Airel) is an ion mobility spectrometer designed to determine the number size distribution of ions in the size range 0.75–45 nm, as well as total (charged and neutral) particles in the size range ~ 2 –45 nm (ref. 51). Previous studies have verified the performance of the NAIS^{52,53}. It consists of two differential mobility analysers (DMAs) in parallel. Each is equipped with 21 electrometers, to separate the mobilities and determine the concentrations of positive and negative ions simultaneously. A corona charger is used when measuring the total particle size distribution.

Particle counters. Several particle counters with different 50% cut-offs were deployed at the CLOUD chamber including two DEG-CPCs^{54,55} (1.5 and 2.7 nm cut-off), one butanol CPC (TSI 3776, 3.2 nm cut-off) and one Particle Size Magnifier (PSM, Airmodus, model A10)⁵⁶. The PSM was run in scanning mode and was used to determine the number size distributions between 1.4 nm and 3.4 nm mobility diameter.

Volatility of oxygenated organics. Recent studies have focused on the formation mechanism of highly oxygenated organics^{17,57,58}. Here we considered the propagation and termination reactions as proposed in refs 57 and 59. We used the radicals from α -pinene ozonolysis proposed in ref. 60 as a starting point and evaluated the possible chemical structures for monomers and dimers (Extended Data Fig. 3). We assume that dimers are covalently bound^{15,17}. This is supported by the chemical formulae of the observed compounds which cannot be explained by a cluster consisting of two monomers.

Instead of assuming an average reduction of the saturation vapour concentration with oxidation, we used this set of chemical structures to calculate the saturation vapour concentration with SIMPOL²⁰.

We then plotted the oxygen to carbon ratio (O:C) as a function of C^* (see Extended Data Fig. 4). We applied a linear least squares fit and used the fit parameters to estimate the volatility for molecules for which we did not derive the structure. The intermediate cluster volatilities were roughly estimated assuming different numbers and types of functional groups (aldehydes, ketones, hydroperoxyacids). The concentration of these clusters is low and will therefore not influence the results significantly. SIMPOL was originally derived at 293 K, but a temperature dependence is given. Thus, we extrapolated C^* to 278 K (resulting in approximately one order of magnitude lower C^* values). Then we separated all observed HOM peaks into volatility regimes¹⁸, as shown in Fig. 2a and b. For this, the HOM concentrations observed in CLOUD for a steady-state run (1209) with ~ 600 p.p.t.v. of injected α -pinene was used. It needs to be noted that the SIMPOL data set does not contain the smallest saturation vapour pressures (as they are difficult to measure quantitatively). Thus, the predicted saturation vapour concentrations for low-volatility compounds could deviate from the actual values. However, the binned volatility distribution is rather flat especially in the ELVOC range. So even if the saturation concentration were to deviate by an order of magnitude, this would not change the conclusions of this work.

Aerosol growth model. The net condensation flux is defined as⁶¹:

$$\dot{m}_{i,p}^c = N_p \frac{(\pi/4(D_p + D_i)^2)}{\alpha_{i,p} \nu_{i,p} \beta_{i,p}} \frac{\dot{m}_{i,p}^c}{\alpha_{i,p} \nu_{i,p} \beta_{i,p}} = \alpha_{i,p} \nu_{i,p} C_i^0 [S_i - a'_{i,p}] \quad (4)$$

deposition rate
of vapours at the surface
particle–vapour
collision cross-section
 $\alpha_{i,p}$
driving force
of condensation
 $F_{i,p}$

with N_p the particle number concentration, D_p the particle diameter, D_i the vapour diameter, $\alpha_{i,p}$ the accommodation coefficient, the vapour concentration C_i^0 and the saturation vapour concentration of C_i^0 . In the following the indicated terms of equation (4) will be further explained.

Deposition rate coefficient. In the molecular regime the collision cross-section is the appropriate metric of a collision probability. Here we assume hard-sphere limit, neglecting charge interactions. The deposition rate coefficient is corrected for the transition regime using the $\beta_{i,p}$ correction factor, to account for non-continuum effects, that is⁶²:

$$\beta_{i,p} = \frac{\text{Kn}(1 + \text{Kn})}{\text{Kn}^2 + \text{Kn} + 0.283\text{Kn}\alpha_{i,p} + 0.75\alpha_{i,p}}; \quad \text{Kn} = \frac{2\lambda}{D_p} \quad (5)$$

The $\beta_{i,p}$ correction term and the mass accommodation coefficient $\alpha_{i,p}$ are connected, as the correction term considers the onset of the gas-phase concentration gradients near the particle. For very small particles ($\text{Kn} \gg 1$), no gradients exist. However, for very large particles ($\text{Kn} \ll 1$), the gas concentration at the particle surface can be near zero even with $\alpha_{i,p} < 1$. The effective mass accommodation coefficient, $a'_{i,p}$, is therefore introduced as well.

For the collision between vapours and ultrafine particles, the reduced mass $\mu_{i,p}$ needs to be considered; $v_{i,p}$ is then the centre of mass velocity:

$$v_{i,p} = \sqrt{8RT/(\pi\mu_{i,p})}; \quad \mu_{i,p} = M_i M_p / (M_i + M_p) \quad (6)$$

The two first terms—collision cross-section and the deposition rate—can be combined. Instead of using the cross-section, the suspended surface area ($N_p \pi D_p^2$) can be used. The modified deposition rate coefficient is then given by:

$$s_{i,p} = \frac{(D_p + D_i)^2}{D_p^2} \frac{v_i}{4} \alpha_{i,p} \beta_{i,p} \quad (7)$$

Condensation sink. Combining the surface area and the deposition rate coefficient we can calculate the collision frequency, which is the frequency with which species i collides with the particle surface:

$$\nu_{i,p}^c = v_i \beta_{i,p} (\alpha_{i,p} = 1) (\pi D_p^2 N_p) \quad (8)$$

The condensation sink, $k_c = \sum_p \alpha_{i,p} \nu_{i,p}^c$, gives the actual time constant for interaction of vapours with particles. The condensation sink is also the fundamental equilibration timescale between the gas and particle phases when condensation is the main loss of vapours.

Driving force of condensation. The driving force of condensation $F_{i,p}$ and excess saturation ratio $S_{i,p}^{XS}$ are:

$$F_{i,p} = [C_i^v - a'_{i,p} C_i^0] = C_i^0 \underbrace{[S_i - a'_{i,p}]}_{S_{i,p}^{XS}} \quad (9)$$

The saturation ratio (gas-phase activity) is $S_i = C_i^v / C_i^0$. The term $a_{i,p}$ is the activity of the species i at the condensed-phase surface of the particle ($a_{i,p} = X_{i,p} \gamma_{i,p}$, Raoult term), where $X_{i,p} = C_{i,p}^s / C_p^s$ is the mass fraction, and $\gamma_{i,p}$ the mass based activity coefficient in the organic condensed phase. Owing to their curved surfaces, the activity of a small particle— $a'_{i,p} = a_{i,p} K_{i,p}$ —includes the Kelvin term $K_{i,p}$. The Kelvin term is defined as⁶¹:

$$K_{i,p} = 10^{D_{K10}/D_p} = \exp\left(\frac{4\sigma_i M_i}{RT\rho_i D_p}\right) \quad (10)$$

$$D_{K10} = \log_{10}(e) \times \frac{4\sigma_i M_i}{RT\rho_i} \quad (11)$$

with the surface tension σ , the molar weight M and the density ρ . For very small particles a large supersaturation is needed to allow for condensation. For $\sigma = 0.023 \text{ N m}^{-1}$, a molar weight of 300 g mol^{-1} at 300 K , $D_K = 3.75 \text{ nm}$. Any charge effect on the growth rate would appear in either an enhancement to the collision cross-section, $\sigma_{i,p}$, due to charge-dipole interactions, or a change in the effective Kelvin diameter reflecting enhanced stability of small clusters. Further investigation of a possible enhancement in the growth rate caused by ions requires dedicated experiments.

Equilibrium solution. At equilibrium, $F_{i,p}$ is zero. In this case, equilibrium partitioning is the basis for organic aerosol calculations. Aerosol partitioning theory describes the condensation and evaporation of gas phase species on or from an aerosol surface⁶³. The fraction of the condensed phase (s) of a species i in the suspended aerosol particle within the partitioning framework is defined as:

$$f_i^s = \frac{1}{1 + C_i^s / C_{OA}^s} \quad (12)$$

C_i^s is the effective saturation concentration of the vapour and C_{OA}^s the concentration of species k in the particle phase.

Steady-state solution. Organic aerosol production, P_i , (or loss) is inherently not an equilibrium process, but many terms will reach a steady state in different situations. There are two relevant limits: one where condensation to suspended particles controls the vapour concentrations on a timescale given by the condensation sink ($\alpha'_{i,p} \nu_{i,p}^c$), and one where losses, k_i (that is, wall losses), control those vapour concentrations. We are interested in the steady-state saturation ratios S_i^{ss} and excess saturation ratio $S_{i,p}^{XS,ss}$.

When losses control the steady-state, $S_i^{ss} = (P_i / C_i^0) / k_i$. If the suspended particles control the steady-state, the excess saturation ratio will be in steady state. A fraction of P_i will go to vapours and a fraction to the particles. The latter fraction will be approximately f_i^s .

$$\phi_i^{XS} = f_i^s P_i = \alpha' k^c C_i^0 [S_i - a'_i] = \alpha' k^c C_i^0 S_{i,p}^{XS,ss}; \quad S_{i,p}^{XS,ss} = f_i^s \frac{P_i / C_i^0}{\alpha' k^c} \quad (13)$$

$S_{i,p}^{XS,ss}$ is a key diagnostic for organic condensation. If $S_{i,p}^{XS,ss} \gg 1$, the condensation will be essentially 'non-volatile' ($a'_{i,p}$ will have no influence on the condensation), while if $S_{i,p}^{XS,ss} \leq 1$ then the condensation will be 'semi-volatile'. Finally, if $S_{i,p}^{XS,ss} \ll 1$, species i cannot be an important driver of the condensation, as $a'_{i,p}$ cannot grow larger than S_i during net gas-phase production.

Dynamic volatility-distribution modelling of aerosol growth. From ref. 15, where the yields were derived from the same experiments, we know the molar yield of HOMs to be roughly $\sim 2.9\%$ from α -pinene ozonolysis. The molar weight of the HOMs is on average twice the molar weight of α -pinene, and we approximate a mass yield of the HOMs of about 6% . The HOMs used include monomers, dimers and intermediate compounds as seen by the nitrate-CI-API-TOF. The concentration of other neutral multimers was either too low or below detection limit (and thus also too low) to contribute significantly to the growth and were neglected in the model. The dynamic volatility-distribution model then condenses the observations into nine volatility bins ranging from $C^* = 10^{-8} \mu\text{g m}^{-3}$ to $C^* = 1 \mu\text{g m}^{-3}$. ELVOC and LVOC were defined as $C^* < 10^{-4.5} \mu\text{g m}^{-3}$ and $10^{-4.5} \mu\text{g m}^{-3} < C^* < 10^{-0.5} \mu\text{g m}^{-3}$ respectively, which is slightly modified compared to ref. 18. This is justified as species with $C^* = 10^{-4} \mu\text{g m}^{-3}$ (typically ELVOC) behaved rather like LVOC, that is, the condensation flux increases with diameter. In Fig. 1 we have seen that the measured HOMs alone cannot explain the observed growth in all size ranges. Therefore, a larger yield of $C^* = 1 \mu\text{g m}^{-3}$ was assumed (light shaded area in Extended Data Fig. 5a), which represents the compounds participating in the formation of the traditional secondary organic aerosol (SOA). Species with $C^* \leq 10^{-8} \mu\text{g m}^{-3}$ were brought into one single bin with $C^* = 10^{-8} \mu\text{g m}^{-3}$. The CI-API-TOF transmission calibration was multiplied by a factor of 1.3, which is within the transmission efficiency uncertainties. The resulting HOM distribution (in percentage) is displayed in Extended Data Fig. 5a.

Using this adjusted HOM distribution, we modelled the growth rate due to condensation assuming no Kelvin effect. Extended Data Fig. 6 shows that the model overestimates the early growth rate and substantially underestimates the observed particle growth rates at larger sizes (blue dashed line). In a next step we modified the charging efficiencies, to match the observation better. Our best result was achieved with values of $[0.5, 0.4, 0.3, 0.1]$ for the VBS bins from 10^{-4} to $10^{-1} \mu\text{g m}^{-3}$, meaning that we increased the raw measured values by $[2, 2.5, 3.3, \text{and } 10]$. Still, it is not possible to describe the observations as depicted by the solid blue line in Extended Data Fig. 6.

Therefore it is essential to introduce the Kelvin effect to reproduce the observed growth rate. In the model we use a Kelvin diameter $D_K = 3.75 \text{ nm}$. This corresponds to a surface tension of 23 mN m^{-1} , which is a reasonable value for organics⁶⁴. If we attempt to model the observed growth using the HOM volatility distribution in Extended Data Fig. 5a, Extended Data Fig. 6 shows that the model substantially underestimates the observed particle growth (pink dashed line), as expected.

The efficiency of HOM charging by the nitrate anion (NO_3^-) depends upon the number and location of OOH groups²¹. As the probability of two OOH groups at optimal configuration is highest for the least volatile species (ELVOC), their charging efficiency is near unity. For products with higher volatility (LVOC) the efficiency decreases. Many of the oxidized monomers might still have a stiff carbon 4-ring backbone hindering an optimal cluster formation between two OOH groups and the nitrate ion. This decreased charging efficiency has yet to be experimentally quantified. Cycloalkene experiments indicate that the nitrate-CI-API-TOF indeed underestimates the low-oxygenated compounds, if compared with the acetate-CIMS⁶⁵, while the concentration for highly oxygenated compounds is similar. The ELVOC bins cannot be changed to a great extent as this would yield an overestimation in the growth rate at sizes below 3 nm .

Adjusting both the LVOC concentrations and the Kelvin term, it is possible to explain the observed size dependent behaviour in Fig. 3. Our best fit was achieved with charging efficiencies of $[0.5, 0.25, 0.1, 0.1]$ for the VBS bins from 10^{-4} to $10^{-1} \mu\text{g m}^{-3}$ and a Kelvin diameter $D_K = 3.75 \text{ nm}$. The final adjusted yields can be seen in Extended Data Fig. 5b, which displays the HOM fraction in the corresponding volatility bins (in percentage). Other tested Kelvin diameters (for example, $D_K = 4.5 \text{ nm}$) yielded a slightly worse agreement with the measurements, the qualitative picture, however, remained the same. Increasing D_K requires an additional adjustment of the ELVOCs to match the observations, so that several parameter combinations will yield similar results. However, very large D_K are very unlikely, as there is not much space to increase the ELVOC concentration due to the nitrate-CI-API-TOF measurement principle.

Here we do not attempt to constrain the volatility distribution exactly. We show that the distribution matters in the formation of particles. ELVOC condensation dominates the growth up to ~ 1.5 nm. Beyond this size, LVOC can contribute and drive the growth. It should be noted that the HOM distribution will change with chamber operating conditions (temperature, α -pinene concentration, particle concentration).

Here we only show two representative runs, but very different cases. We did not perform experiments with pre-existing particles in the chamber, at least not in such an amount to overcome the sink due to the wall ($k_{\text{wall}} \approx 10^{-3} \text{ s}^{-1}$ versus $k_{\text{cond}} \approx 10^{-4} \text{ s}^{-1}$ or lower). The wall loss does in some way simulate the sink due to pre-existing particles. The measured gas-phase concentration is a result of the existing sink and source terms. These terms will be somewhat different in the chamber compared to ambient conditions. Thus, we cannot say that under the same α -pinene and ozone concentrations the growth is the same. But, measuring the same volatility distribution of HOMs in the ambient (and at the same temperature) should yield similar results. The exact evolution of the particle size and the contribution of the volatility bins will always depend on the observed volatility distribution of the HOM species. The volatility distribution itself will depend on the temperature and the oxidants (for example, NO_x will hinder the formation of ELVOC, lowering the yield¹⁷). But the approach proposed here and the corresponding conclusion will still be applicable.

Model details. For the simulations we assumed a mono-disperse population of nucleated particles at an initial size of 1.2 nm mobility diameter or 0.9 nm physical diameter (which is approximately the monomer size). The key parameter is the concentration gradient (see equation (9)), which in turn reflects the differences in activity between the gas phase (the saturation ratio) and the particle phase (here the mass fraction). This can be seen in Extended Data Fig. 7a. The gas phase is characterized by the balance between the production rate of the α -pinene oxidation products and wall losses yielding a stable gas-phase saturation ratio. In contrast, the condensed phase activities drop as soon as the particles grow and the Kelvin effect decreases.

Looking at the excess saturation (Extended Data Fig. 7b), the least volatile species (mostly ELVOC) have a significant excess saturation at all times; the condensed phase activity is always much lower than the gas-phase saturation ratio. The more volatile species are near equilibrium at the beginning, only gradually (if ever) developing a significant driving force of condensation. The most volatile species are in equilibrium all of the time with a diminishing mass fraction in the condensed phase. For < 2.5 nm, the particles are unstable, with the majority of their constituents showing activities $\gg 1$. They can only grow as a consequence of the excess saturation ratio of the ELVOCs. If the production were rapidly stopped, the particles would evaporate. Extended Data Fig. 7b also shows the condensed phase mass fraction and thus the chemical composition of the particle. Particles < 2.5 nm are mainly composed of ELVOC dominated by species with $C^* = 10^{-8} \mu\text{g m}^{-3}$. For larger particles the LVOC mass fraction increases until each contributes equally to the particle composition. The two most volatile bins never contribute substantially to the particle composition as their gas-phase saturation ratio is too low.

Extended Data Fig. 7c shows the absolute driving force of condensation and the equilibrium concentration of the different volatile species over the growing particles. Here, this transition from ELVOC to LVOC dominated growth is evident in the driving force of condensation. Owing to the stiff coupled differential equations tight tolerances on the solver are required for the solution to converge accurately. **Appearance times and growth rate estimation of clusters and aerosols.** The appearance times of clusters and aerosols allow us to investigate the growth process. Cluster and particle appearance times, defined as the 50% rise time of the concentration of a cluster or size channel⁶⁶, were derived for APi-TOF, PSM, NAIS, DEG-CPCs, nRDMA and nano-SMPS. The corresponding diameters (leading edge diameter) were then plotted against the time. The temporal evolution is then representing the growth rate. For linear evolution, a linear fit was applied; the slope yields the growth rate. Extended Data Fig. 1 combines all the calculated appearance times for one example run. It shows an excellent agreement between the different methods and instruments.

To determine the appearance time for APi-TOF, NAIS, and PSM, concentrations in each size bin were analysed and the time when the concentration reaches 50% of its maximum value after the start of a nucleation experiment was determined and linked with the diameter midpoint of the size bin. The growth rate was obtained from a linear fit of the appearance times and the corresponding diameters. For the PSM the growth rate could be determined for the size range 1.5–3.2 nm. For the NAIS: (1) 1.4–3 nm, (2) 5–15 nm and (3) 15–30 nm. In the APi-TOF, appearance times of the monomers, dimers, trimers and tetramers were determined.

A normal (Gauss) function was applied to the size distribution data^{2,67}. The position of the full-width at half-maximum (FWHM) was then defined as the 50% rise time. Nano-SMPS growth rates were determined for the following size

ranges: (1) 5–15 nm, (2) 15–30 nm, (3) 30–60 nm and (4) > 60 nm. In these size ranges, a constant growth rate for constant HOM concentration was observed, so we did not further differentiate these ranges in Fig. 1. For the nRDMA: (1) 1.1–3 nm, (2) 2–7 nm.

The DEG-CPC method was slightly different. In previous studies⁶, the 1% threshold of the CPC and the initial rise of the concentration was used to further extend the growth rate analysis to lower diameters. We decided to also use this approach for the DEG-CPCs. However, owing to the high noise, it was often difficult to determine the 1% rise time, thus the 5% rise time of the DEG-CPCs was used instead, yielding similar results.

Growth rate uncertainties. The method uncertainty is estimated⁶⁶ to be approximately 50%. To consider the run-to-run uncertainty, we used σ_{fit} as retrieved from the linear fit uncertainty to determine the growth rate (GR). The overall uncertainty then scales as follows:

$$\sigma_{\text{tot}} = \sqrt{0.25[\text{GR}]^2 + \sigma_{\text{fit}}^2} \quad (14)$$

The growth rates in Fig. 1c, d correlate reasonably well with the HOM concentration. Growth rates of larger sizes correlate with a Pearson's correlation coefficient of 0.94, growth rates at smaller size with a Pearson's correlation coefficient of 0.7. The lower correlation at the smaller sizes can be explained by the higher measurement uncertainty at these size ranges, compared to larger sizes.

Parameterization of first steps of growth and global aerosol modelling. We are especially interested in the first steps of growth, that is, from the nucleated cluster size to 3 nm, as there the coagulation losses are highest. In the global model we use here⁶⁸, nucleated clusters have a diameter of 1.7 nm, and particles must grow to 3 nm before being advected through the atmosphere in the nucleation mode. Therefore we parameterize the growth rate in the size range 1.7–3 nm. We use the size-resolved growth rates from the HOM volatility-distribution modelling results to derive a size-dependent parameterization. The Kelvin effect increases the growth rate with increasing size. The considered size range (1.7–3 nm) is small enough that we can approximate the dependence on the particle diameter D_p as linear. We thus parameterize the growth rate (in nm h^{-1}) by fitting the two-dimensional function ($[\text{HOM}]$ in cm^{-3} , D_p in nm):

$$\text{GR} = kD_p[\text{HOM}]^p \quad (15)$$

to the HOM volatility-distribution modelling results, with the free parameters $k = (5.2 \pm 0.4) \times 10^{-11}$ and $p = 1.424 \pm 0.004$. Here the uncertainties are those from the fit only; they reflect how well the function describes the data but do not represent the full uncertainty in the parameterization. The parameterization is intended to describe the size-dependent growth that we observe, and does not necessarily reflect the underlying mechanism. Therefore, extrapolations to very high values ($> 5 \times 10^8 \text{ cm}^{-3}$) and low values ($< 2 \times 10^6 \text{ cm}^{-3}$) may not be reliable, as it is likely that the parameterized growth rates deviate from the true growth rates. Such high biogenic HOM values, however, are not expected in the field and should not impact the global modelling results. Conversely, low HOM concentrations far below $2 \times 10^6 \text{ cm}^{-3}$ are expected far from sources of terpenes, especially over oceans and the upper free troposphere. From Fig. 1 it is evident that the growth rate at $[\text{HOM}] < 2 \times 10^6 \text{ cm}^{-3}$ is $< 1 \text{ nm h}^{-1}$. Under these conditions, growth is driven by condensation of sulfuric acid, and uncertainties in the parameterization of the very small organic contribution are not expected to affect the results significantly.

This parameterization provides a refined estimate of the growth rate between 1.7 and 3 nm, which is appropriate for models of atmospheric aerosol that treat SOA condensation kinetically. To implement the parameterization, a mechanism and yield for the production of HOMs is required. In our model, HOMs are simulated as being produced directly from the oxidation of monoterpenes (MT) and lost to the condensation sink (CS) in a steady-state approximation:

$$[\text{HOM}] = (Y_1 k_1 [\text{MT}][\text{O}_3] + Y_2 k_2 [\text{MT}][\text{OH}]) / \text{CS} \quad (16)$$

where Y_1 , the yield of HOMs from the ozonolysis of monoterpenes, is 2.9%, and Y_2 , the yield from the OH-oxidation, is 1.2%. The yields were determined from the nitrate-CI-APi-TOF and PTR-TOF measurements in the CLOUD chamber¹⁵. The constants k_1 and k_2 are the temperature dependent reaction rate constants of α -pinene with ozone and hydroxyl radicals, respectively⁶⁹. Thus the numerator of equation (16) represents the production of HOMs and the denominator the losses.

We do not quote a similar parameterization for growth rates at larger sizes, because it is clear that the nitrate-CI-APi-TOF does not see all of the more volatile molecules that condense onto larger particles, many more compounds are likely to participate than those present in the CLOUD chamber, and at these larger sizes the kinetic condensation approach should be complemented by an equilibrium partitioning treatment (for example, ref. 70).

This parameterization represents pure organic growth resulting from biogenic emissions. In the ambient atmosphere, additional organic and inorganic precursors such as sulfuric acid, ammonia, amines and anthropogenic VOCs are also present and influence the growth rate, in addition to the different oxidants. Also temperature and relative humidity could influence the observed growth rates. So, while this parameterization represents a significant advance on the current state of the art, it should not be considered complete. Furthermore, we only consider the size range 1.7 to 3 nm, as the growth in this size range is most decisive for the fate of the freshly nucleated particle⁴.

The parameterization of initial particle growth is incorporated in the global aerosol model GLOMAP-mode⁶⁸, an extension to the TOMCAT chemical transport model⁷¹. GLOMAP includes representations of particle formation, growth via coagulation, condensation and cloud processing, wet and dry deposition and in/below cloud scavenging. The horizontal resolution is 2.8×2.8 degrees and there are 31 vertical sigma-pressure levels extending from ground level to 10 hPa. Aerosol in the model is formed of four components: black carbon, organic carbon, sea salt and sulfate, and is advected through the atmosphere in seven log-normal size modes. These are hygroscopic nucleation, Aitken, accumulation and coarse modes, and non-hygroscopic Aitken, accumulation and coarse modes. Formation of secondary particles in the model is based on CLOUD measurements of ternary H₂SO₄-organic-H₂O nucleation detailed in ref. 25 and on a parameterization of binary H₂SO₄-H₂O nucleation⁷². Simulations are run for the year 2008.

In the aerosol model, particles grow by irreversible condensation of monoterpene oxidation products and sulfuric acid. Monoterpene emissions in the model are taken from the database of ref. 73. Our measurements¹⁵ provide HOM yields of 2.9% from the oxidation of α -pinene by ozone and 1.2% from the hydroxyl radical. In ref. 58 a substantially higher HOM yield was observed from endocyclic monoterpenes such as α -pinene than from exocyclic monoterpenes. These two types are roughly equally abundant in the atmosphere. Thus, we account for this by dividing our measured yields by two. In the light of these results, we also divide the organic nucleation rate of ref. 25 by two, since it also assumed all terpenes were represented by α -pinene in the atmosphere. Above 3 nm in diameter, a fixed 13% of the oxidation products of monoterpenes with OH, O₃ and NO₃ (assuming the reaction rates of α -pinene) condense irreversibly onto aerosol particles at the kinetic limit. These oxidized organic molecules are referred to as SORG and are advected through the troposphere as a tracer in the model, while the HOM concentration is calculated assuming a steady state as described earlier. Below 3 nm, organic molecules condense onto particles according to the parameterization, while sulfuric acid molecules condense at the kinetic limit (collision-limited), which is approximately:

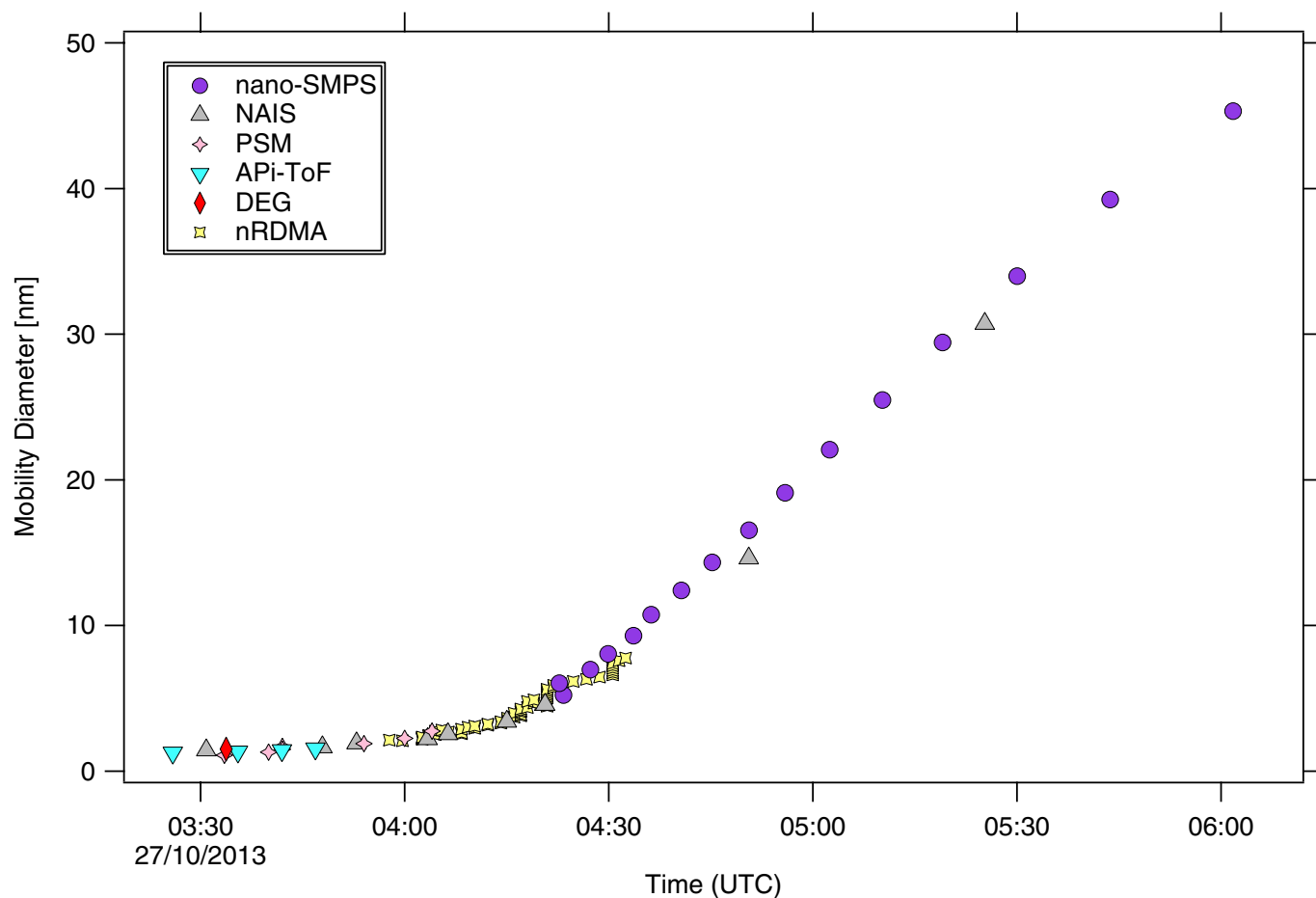
$$GR_S = 7.3 \times 10^{-8} [H_2SO_4] \quad (17)$$

Additional model runs were performed with no organics participating in the initial growth, and with non-volatile size-dependent growth of particles between 1.7 and 3 nm due to condensation of SORG multiplied by the factor determined in ref. 30 for the parameterization of ref. 3,

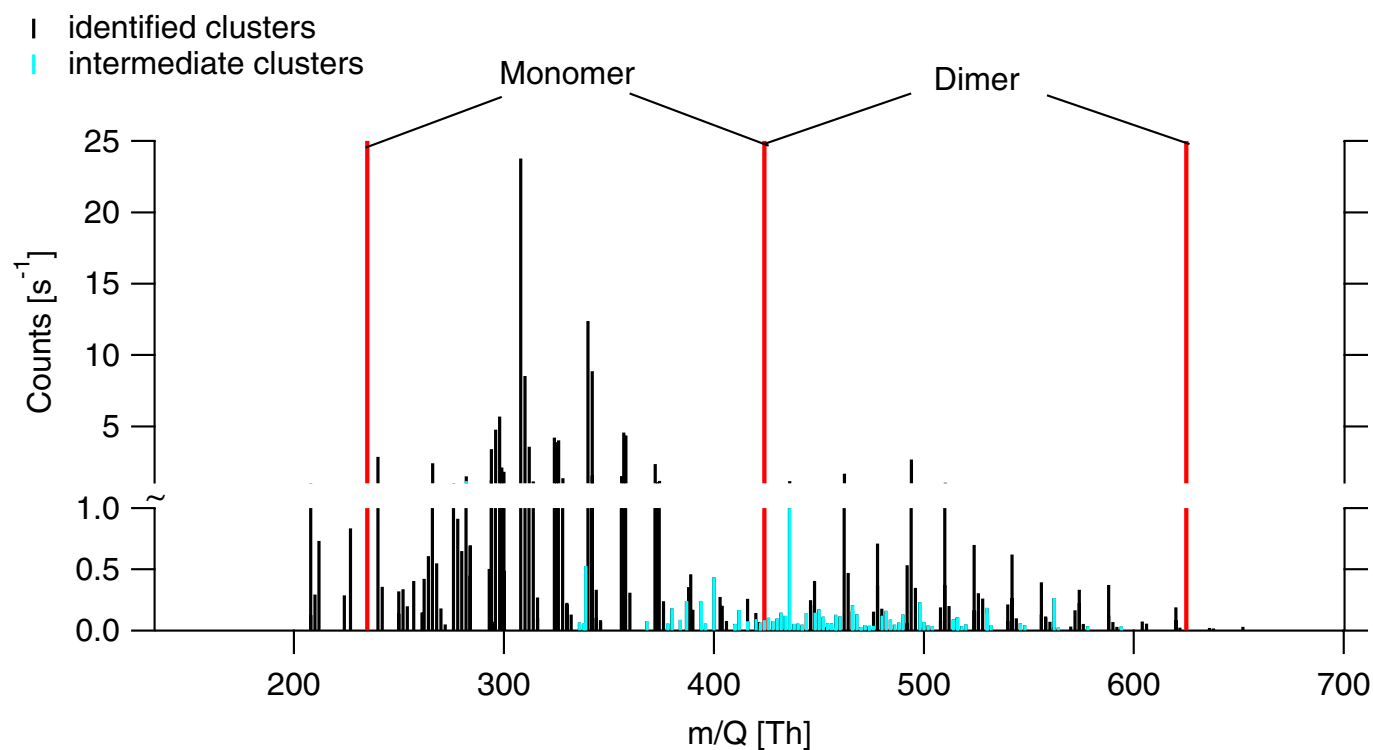
$$k = 0.47D_p - 0.18 \quad (18)$$

where D_p is the particle diameter in nm and the correction is only applied to particles below 2.5 nm. We note that the SORG in GLOMAP is produced with a 13% yield while that in GEOS-chem is produced with a 10% yield. The growth rates in these three cases are shown in Extended Data Fig. 9, together with the HOM concentration in the model.

31. Kirkby, J. *et al.* Role of sulphuric acid, ammonia and galactic cosmic rays in atmospheric aerosol nucleation. *Nature* **476**, 429–433 (2011).
32. Duplissy, J. *et al.* Effect of ions on sulfuric acid-water binary particle formation: 2. Experimental data and comparison with qc-normalized classical nucleation theory. *J. Geophys. Res. Atmos.* **121**, 1752–1775 (2016).
33. Kupc, A. *et al.* A fibre-optic UV system for H₂SO₄ production in aerosol chambers causing minimal thermal effects. *J. Aerosol Sci.* **42**, 532–543 (2011).
34. Duplissy, J. *et al.* Results from the CERN pilot CLOUD experiment. *Atmos. Chem. Phys.* **10**, 1635–1647 (2010).
35. Voigtländer, J., Duplissy, J., Rondo, L., Kürten, A. & Stratmann, F. Numerical simulations of mixing conditions and aerosol dynamics in the CERN CLOUD chamber. *Atmos. Chem. Phys.* **12**, 2205–2214 (2012).
36. Schnitzhofer, R. *et al.* Characterisation of organic contaminants in the CLOUD chamber at CERN. *Atmos. Meas. Tech.* **7**, 2159–2168 (2014).
37. Bianchi, F., Dommen, J., Mathot, S. & Baltensperger, U. On-line determination of ammonia at low pptv mixing ratios in the CLOUD chamber. *Atmos. Meas. Tech.* **5**, 1719–1725 (2012).
38. Jokinen, T. *et al.* Atmospheric sulphuric acid and neutral cluster measurements using CI-API-TOF. *Atmos. Chem. Phys.* **12**, 4117–4125 (2012).
39. Graus, M., Müller, M. & Hansel, A. High resolution PTR-TOF: quantification and formula confirmation of VOC in real time. *J. Am. Soc. Mass Spectrom.* **21**, 1037–1044 (2010).
40. Junninen, H. *et al.* A high-resolution mass spectrometer to measure atmospheric ion composition. *Atmos. Meas. Tech.* **3**, 1039–1053 (2010).
41. Kürten, A., Rondo, L., Ehrhart, S. & Curtius, J. Calibration of a chemical ionization mass spectrometer for the measurement of gaseous sulfuric acid. *J. Phys. Chem. A* **116**, 6375–6386 (2012).
42. Cheng, Y.-S. in *Aerosol Measurement: Principles, Techniques, and Applications* (eds Kulkarni, P. *et al.*) 569–601 (John Wiley & Sons, 2001).
43. Heinritzi, M. *et al.* Characterization of the mass-dependent transmission efficiency of a CIMS. *Atmos. Meas. Tech.* **9**, 1449–1460 (2016).
44. Möhler, O., Reiner, T. H. & Arnold, F. The formation of SO₅ by gas phase ion-molecule reactions. *J. Chem. Phys.* **97**, 8233–8239 (1992).
45. Kürten, A., Rondo, L., Ehrhart, S. & Curtius, J. Performance of a corona ion source for measurement of sulfuric acid by chemical ionization mass spectrometry. *Atmos. Meas. Tech.* **4**, 437–443 (2011).
46. Brunelli, N. A., Flagan, R. C. & Giapis, K. P. Radial differential mobility analyzer for one nanometer particle classification. *Aerosol Sci. Technol.* **43**, 53–59 (2009).
47. Wang, J., McNeill, V. F., Collins, D. R. & Flagan, R. C. Fast mixing condensation nucleus counter: application to rapid scanning differential mobility analyzer measurements. *Aerosol Sci. Technol.* **36**, 678–689 (2002).
48. Jiang, J. *et al.* Transfer functions and penetrations of five differential mobility analyzers for sub-2 nm particle classification. *Aerosol Sci. Technol.* **45**, 480–492 (2011).
49. Wang, S. C. & Flagan, R. C. Scanning electrical mobility spectrometer. *Aerosol Sci. Technol.* **13**, 230–240 (1990).
50. Kulkarni, P., Baron, P. A. & Willeke, K. *Aerosol Measurement: Principles, Techniques, and Applications* (John Wiley & Sons, 2011).
51. Mirme, S. & Mirme, A. The mathematical principles and design of the NAIS—a spectrometer for the measurement of cluster ion and nanometer aerosol size distributions. *Atmos. Meas. Tech.* **6**, 1061–1071 (2013).
52. Asmi, E. *et al.* Results of the first air ion spectrometer calibration and intercomparison workshop. *Atmos. Chem. Phys.* **9**, 141–154 (2009).
53. Gagné, S. *et al.* Intercomparison of air ion spectrometers: an evaluation of results in varying conditions. *Atmos. Meas. Tech.* **4**, 805–822 (2011).
54. Wimmer, D. *et al.* Performance of diethylene glycol-based particle counters in the sub-3 nm size range. *Atmos. Meas. Tech.* **6**, 1793–1804 (2013).
55. Iida, K., Stolzenburg, M. R. & McMurry, P. H. Effect of working fluid on sub-2 nm particle detection with a laminar flow ultrafine condensation particle counter. *Aerosol Sci. Technol.* **43**, 81–96 (2009).
56. Vanhanen, J. *et al.* Particle size magnifier for nano-CN detection. *Aerosol Sci. Technol.* **45**, 533–542 (2011).
57. Rissanen, M. P. *et al.* The formation of highly oxygenated multifunctional products in the ozonolysis of cyclohexene. *J. Am. Chem. Soc.* **136**, 15596–15606 (2014).
58. Jokinen, T. *et al.* Rapid autoxidation forms highly oxidized RO₂ radicals in the atmosphere. *Angew. Chem. Int. Ed.* **53**, 14596–14600 (2014).
59. Mentel, T. *et al.* Formation of highly oxidized multifunctional compounds: autoxidation of peroxy radicals formed in the ozonolysis of alkenes deduced from structure product relationships. *Atmos. Chem. Phys.* **15**, 6745–6765 (2015).
60. Zhang, D. & Zhang, R. Ozonolysis of α -pinene and β -pinene: kinetics and mechanism. *J. Chem. Phys.* **122**, 114308 (2005).
61. Seinfeld, J. H. & Pandis, S. N. *Atmospheric Chemistry and Physics: from Air Pollution to Climate Change* (John Wiley & Sons, 2006).
62. Fuchs, N. A. & Sutugin, A. G. *Coagulation rate of Highly Dispersed Aerosols* (Ann Arbor Science, 1970).
63. Pankow, J. F. An absorption model of gas/particle partitioning of organic compounds in the atmosphere. *Atmos. Environ.* **28**, 185–188 (1994).
64. Korosi, G. & Kovats, E. S. Density and surface tension of 83 organic liquids. *J. Chem. Eng. Data* **26**, 323–332 (1981).
65. Berndt, T. *et al.* Gas-phase ozonolysis of cycloalkenes: formation of highly oxidized RO₂ radicals and their reactions with NO, NO₂, SO₂, and other RO₂ radicals. *J. Phys. Chem. A* **119**, 10336–10348 (2015).
66. Lehtipalo, K. *et al.* Methods for determining particle size distribution and growth rates between 1 and 3 nm using the particle size magnifier. *Boreal Environ. Res.* **19**, 215–236 (2014).
67. Kulmala, M. *et al.* Initial steps of aerosol growth. *Atmos. Chem. Phys.* **4**, 2553–2560 (2004).
68. Mann, G. W. *et al.* Description and evaluation of GLOMAP-mode: a modal global aerosol microphysics model for the UKCA composition-climate model. *Geoscientific Model Dev.* **3**, 519–551 (2010).
69. McNaught, A. D. & Wilkinson, A. *Compendium Of Chemical Terminology* Vol. 1669 (Blackwell Science, 1997).
70. Riipinen, I. *et al.* Organic condensation: a vital link connecting aerosol formation to cloud condensation nuclei (CCN) concentrations. *Atmos. Chem. Phys.* **11**, 3865–3878 (2011).
71. Chipperfield, M. P. New version of the TOMCAT/SIMCAT off-line chemical transport model: Intercomparison of stratospheric tracer experiments. *Q. J. R. Meteorol. Soc.* **132**, 1179–1203 (2006).
72. Kulmala, M., Laaksonen, A. & Pirjola, L. Parameterizations for sulfuric acid/water nucleation rates. *J. Geophys. Res. D* **103**, 8301–8307 (1998).
73. Guenther, A. *et al.* A global model of natural volatile organic compound emissions. *J. Geophys. Res. D* **100**, 8873–8892 (1995).
74. Kurtén, T. *et al.* Computational study of hydrogen shifts and ring-opening mechanisms in α -pinene ozonolysis products. *J. Phys. Chem. A* **119**, 11366–11375 (2015).



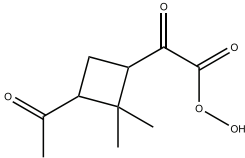
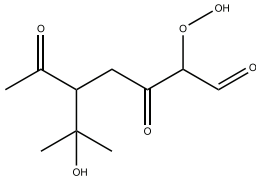
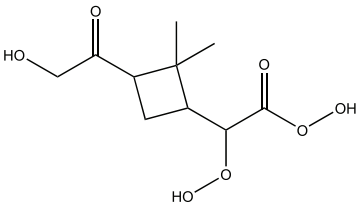
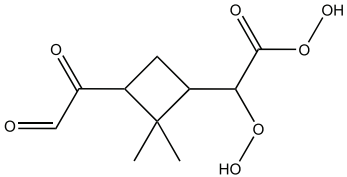
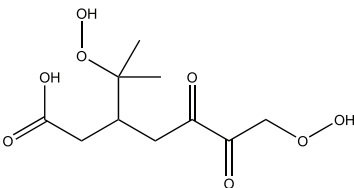
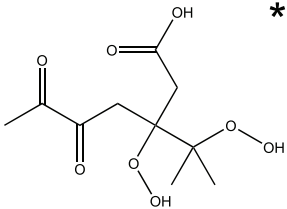
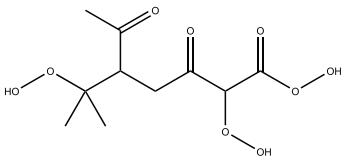
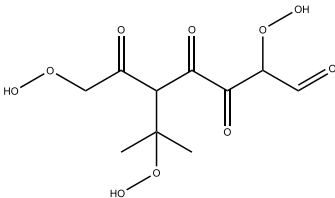
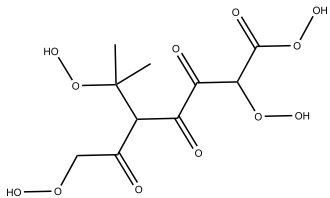
Extended Data Figure 1 | Appearance times of clusters and aerosols as seen by nano-SMPS, NAIS, PSM, APi-TOF, DEG and nRDMA. The different instruments are indicated with different plotting symbols. Instrument descriptions and acronyms can be found in Methods.



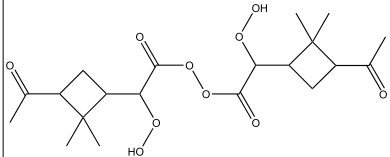
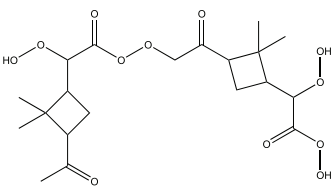
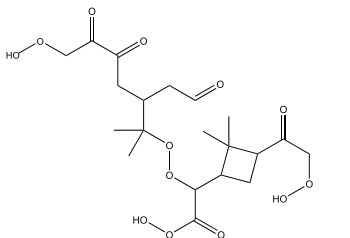
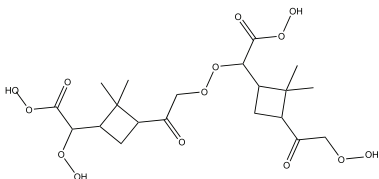
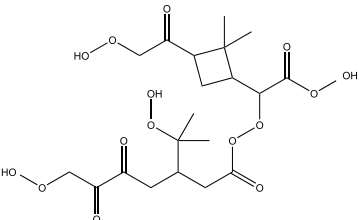
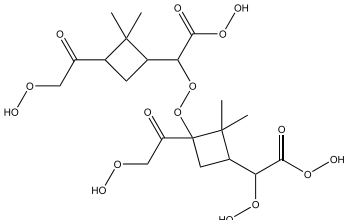
Extended Data Figure 2 | Observed mass spectrum as seen by the nitrate-CI-API-TOF at 278 K and 38% relative humidity. A steady-state mixing ratio of approximately 250 p.p.t.v. of α -pinene was established in the chamber in the presence of 35 p.p.b.v. ozone and no injection of SO_2 . Black bars indicate all identified monomers and dimers, with the red

bars indicating the corresponding m/Q range. Intermediate molecules or clusters (with carbon atoms between 11 and 17) that cannot be explained by the formation mechanism shown in Kirkby *et al.*¹⁵ are indicated by the cyan bars.

a

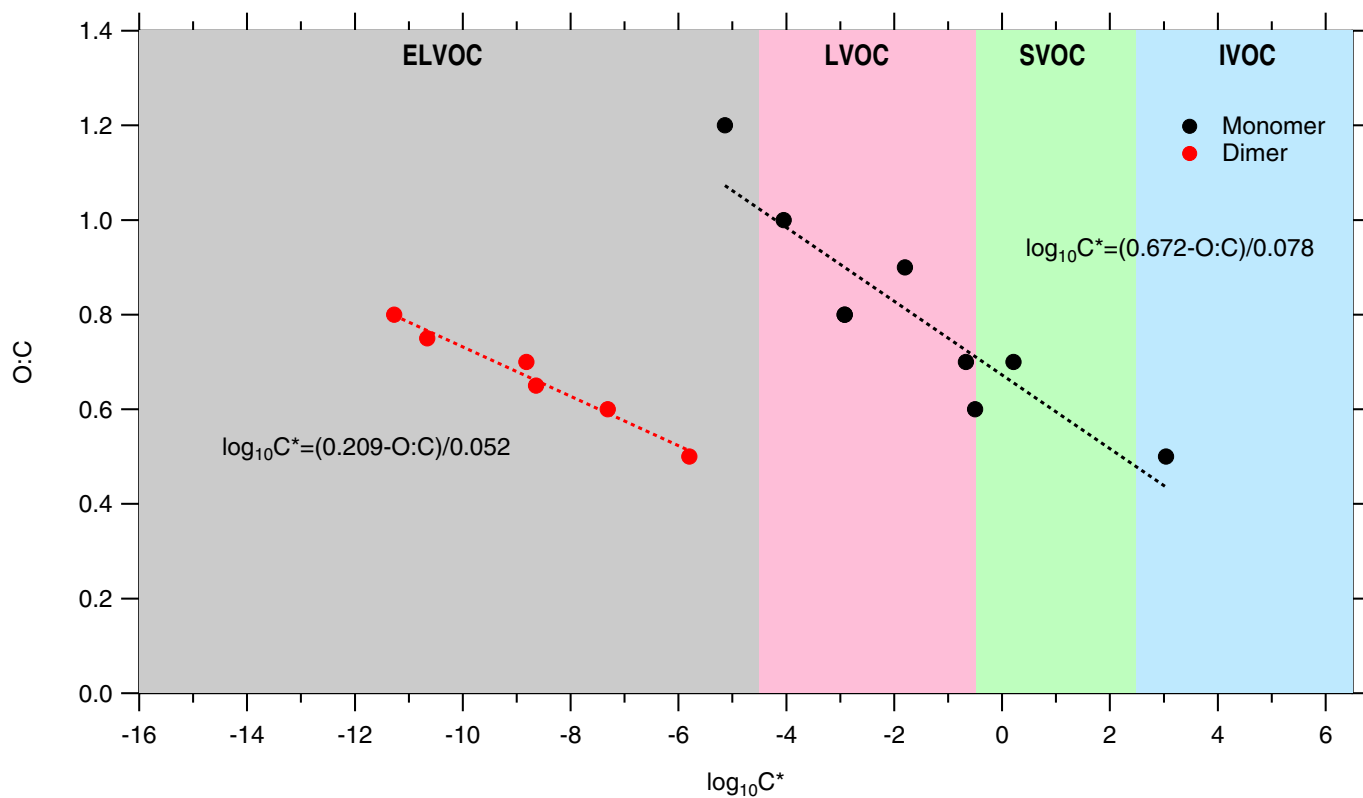
		
Chemical Formula: $C_{10}H_{14}O_5$ $\log C^* = 3.04$	Chemical Formula: $C_{10}H_{16}O_6$ $\log C^* = -0.50$	Chemical Formula: $C_{10}H_{14}O_7$ $\log C^* = -0.67$
		
Chemical Formula: $C_{10}H_{14}O_7$ $\log C^* = 0.21$	Chemical Formula: $C_{10}H_{16}O_8$ $\log C^* = -2.92$	Chemical Formula: $C_{10}H_{16}O_8$ $\log C^* = -2.92$
		
Chemical Formula: $C_{10}H_{16}O_9$ $\log C^* = -1.80$	Chemical Formula: $C_{10}H_{14}O_{10}$ $\log C^* = -4.05$	Chemical Formula: $C_{10}H_{14}O_{12}$ $\log C^* = -5.14$

b

		
Chemical Formula: $C_{20}H_{30}O_{10}$ $\log C^* = -5.8$	Chemical Formula: $C_{20}H_{30}O_{12}$ $\log C^* = -7.31$	Chemical Formula: $C_{20}H_{30}O_{13}$ $\log C^* = -8.64$
		
Chemical Formula: $C_{20}H_{30}O_{14}$ $\log C^* = -8.82$	Chemical Formula: $C_{20}H_{30}O_{15}$ $\log C^* = -10.66$	Chemical Formula: $C_{20}H_{30}O_{16}$ $\log C^* = -11.27$

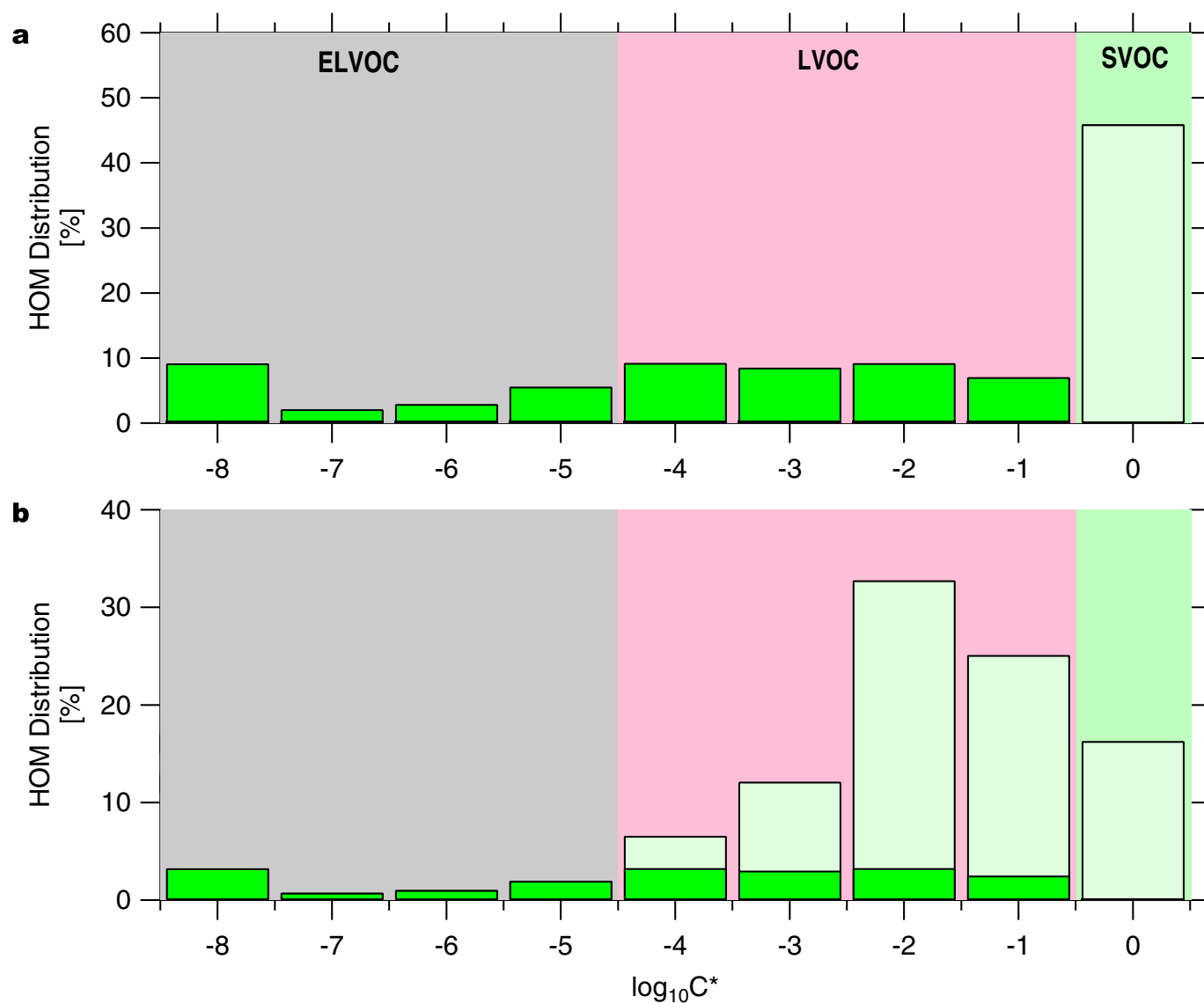
Extended Data Figure 3 | Possible structures of α -pinene oxidation products. **a**, Possible structures of HOM monomer molecules. C^* was estimated using the SIMPOL method (at 293 K). Note that the volatility is less once the ring structure is open. The volatility generally decreases with

increasing oxidation and decreasing temperature. **b**, Possible structures of HOM dimer molecules. C^* was estimated using the SIMPOL method (at 293 K). Structures in boxes with asterisk(s) at the top right corner were confirmed by (*) or taken from (**) Kurtén *et al.*⁷⁴.



Extended Data Figure 4 | Estimation of C^* for monomer and dimer molecules at 293 K. Colours indicate the volatility class based on ref. 18. A linear fit was applied to the C^* estimates (dashed lines). This fit was then applied to all compounds using their O:C ratio to

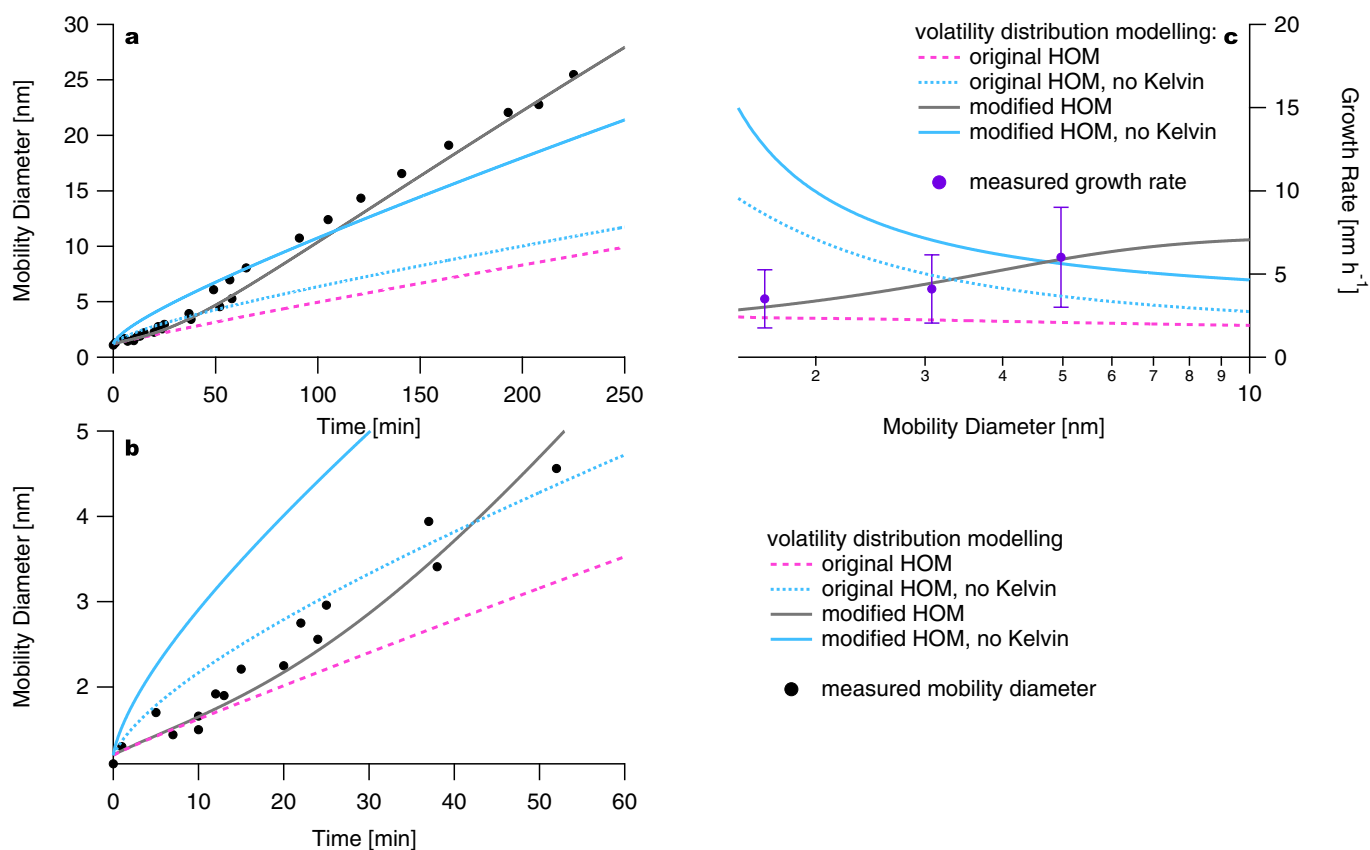
estimate their value of C^* . Volatility bins comprise ELVOC, LVOC, SVOC and IVOC (intermediate volatile organic compounds with C^* from $10^{2.5}$ to $10^{6.5} \mu\text{g m}^{-3}$).



Extended Data Figure 5 | HOM distribution binned to a VBS.

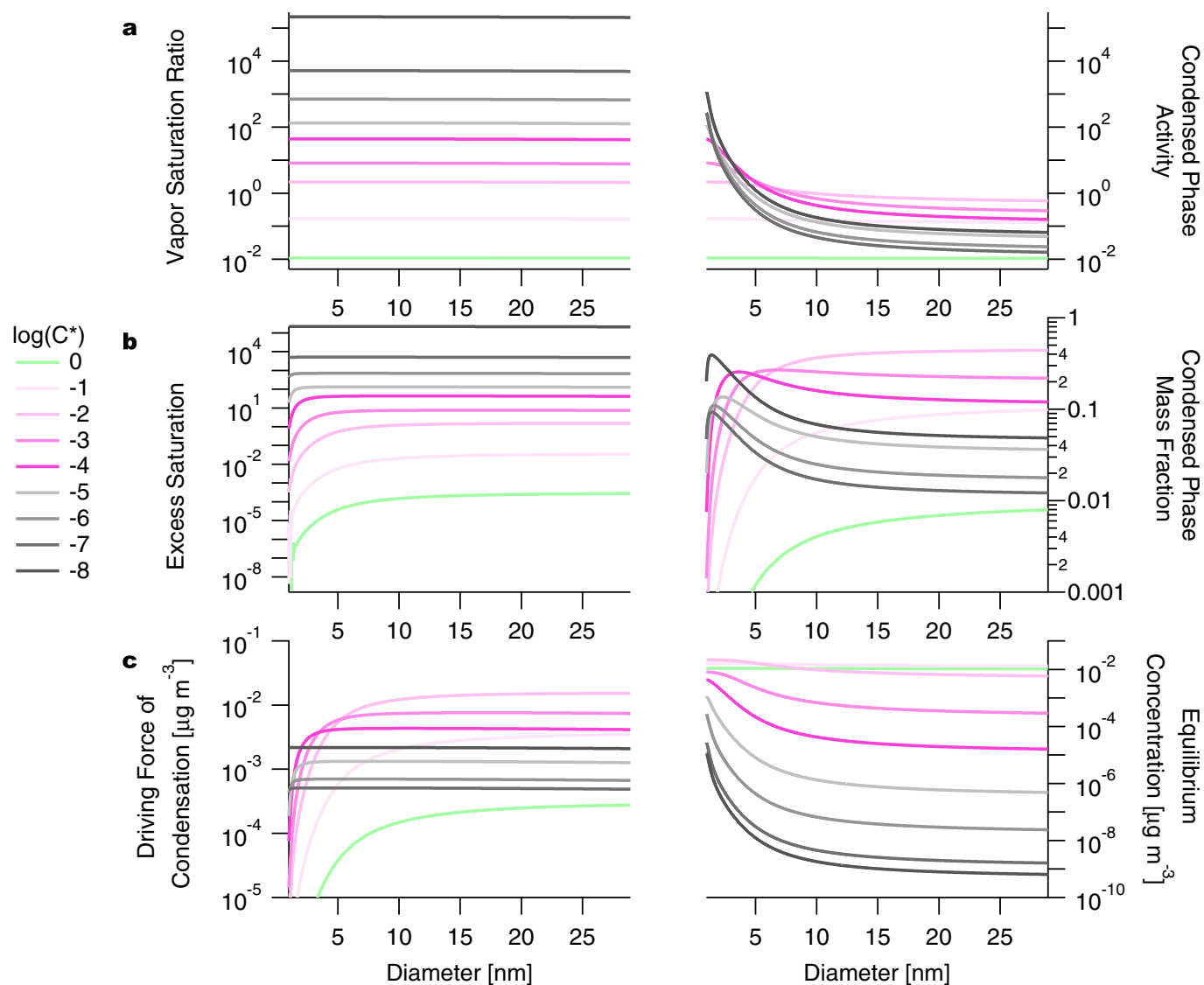
a, Measured HOM distribution (dark green) binned to a VBS. As the nitrate-CI-APi-TOF is expected to underestimate SVOC, which are often observed during secondary aerosol formation in smog chamber studies,

we added a representative SOA bin at $\log C^* = 0$ (light green). **b**, Modified HOM distribution after scaling for the weaker charging efficiency for LVOC (light green). The ELVOC:LVOC:SVOC ratios are **a**, 20:34:46 and **b**, 7:77:16.



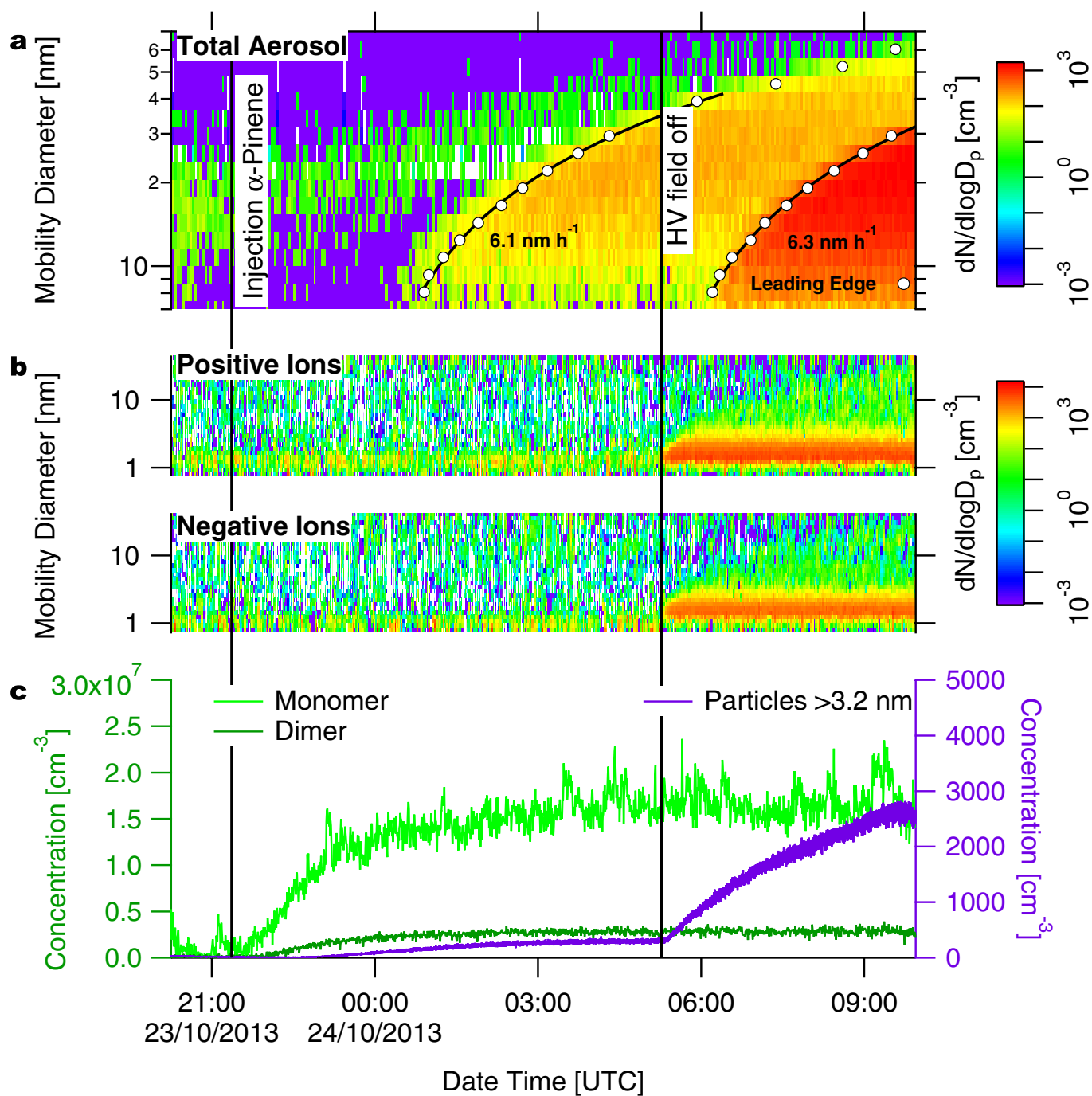
Extended Data Figure 6 | Dynamic volatility-distribution modelling results with and without a Kelvin term and with original and modified HOM volatility distribution for the case of constant HOMs. a, Different model approaches (key at bottom right) compared to the measured diameter evolution. **b,** Enlargement of the first 30 min of the experiment and the first 5 nm of the diameter evolution (key in panel). **c,** Size dependent growth rate for different model approaches (key at bottom right). The Kelvin effect is essential to describe the measured diameter behaviour. Using the original volatility distribution (blue dashed line), the model slightly overestimates the initial growth but strongly underestimates

it at larger sizes. Although considering a Kelvin effect fits the initial growth well, growth at larger sizes is underestimated even more (pink dashed line). By adjusting the HOM volatility distribution in the model with no Kelvin effect, the best fit (blue solid curve) still fails to reproduce the observations, substantially overpredicting growth at small sizes and then underpredicting growth at larger sizes. However, adjusting the volatility distribution and treating the Kelvin effect captures the growth well over the full size range (grey solid line). Error bars indicate the 1σ systematic scale uncertainty of the determined growth rates.



Extended Data Figure 7 | Dynamic volatility-distribution model details. **a**, Vapour (left) and condensed-phase (right) activities during a simulated particle growth event in CLOUD (Fig. 3b, d). Vapours are in steady-state with respect to production and wall loss, with the saturation ratio increasing monotonically with decreasing volatility. **b**, Excess saturation ratios (left) and particle composition (mass fractions; right)

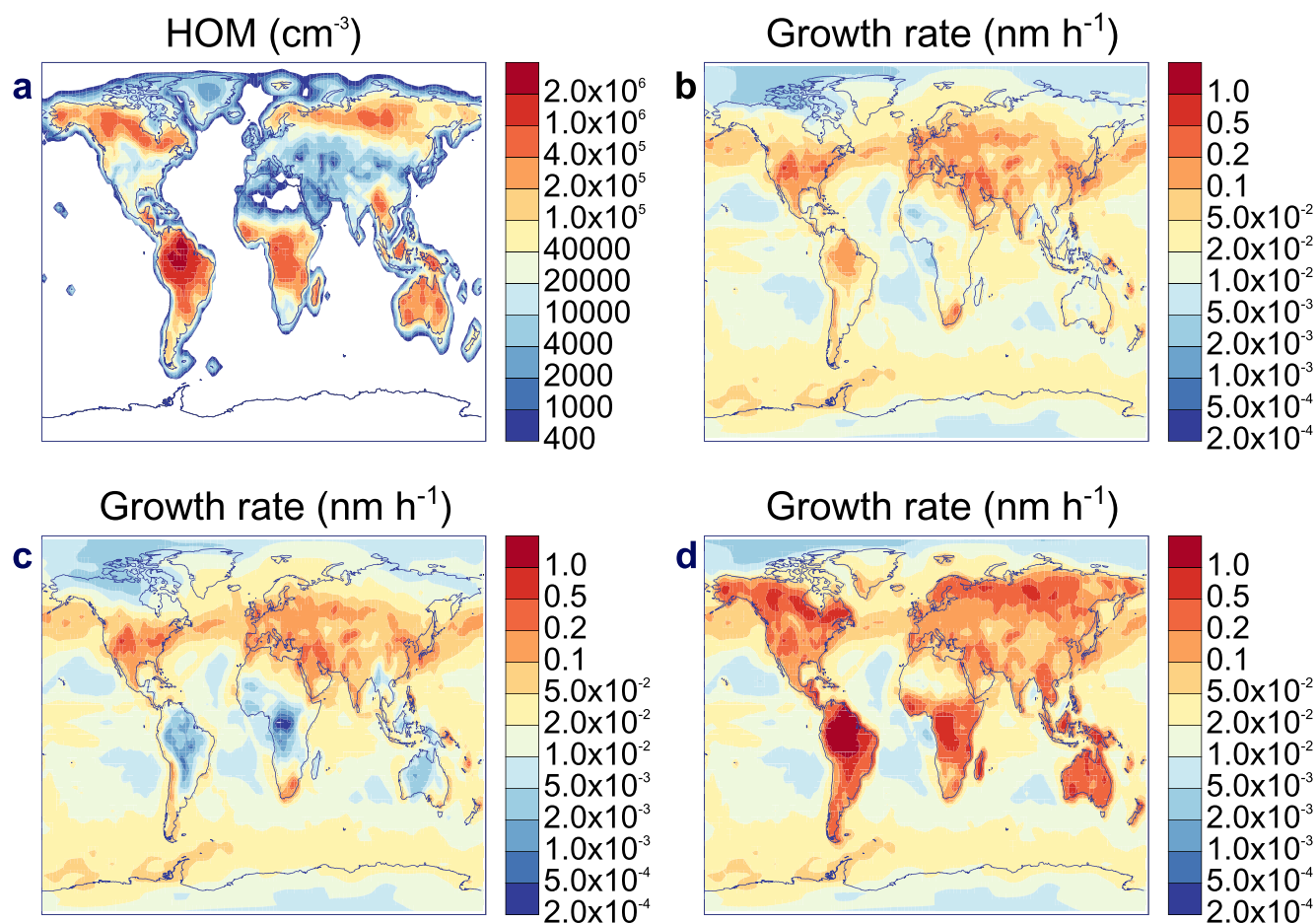
during simulated particle growth event in CLOUD. **c**, Driving force of condensation (left) and equilibrium concentrations of vapours over particles (right) during simulated particle growth events in CLOUD. Volatility is indicated by brightness, with darkest grey corresponding to $C^* = 10^{-8} \mu\text{g m}^{-3}$ (see key).



Extended Data Figure 8 | Typical experiment during CLOUD8.

α -Pinene was injected under neutral conditions. Once stable conditions were achieved, the clearing (HV) field was turned off allowing ions from Galactic cosmic rays to remain in the chamber. This immediately yields a second nucleation. **a**, The nano-SMPS size distribution; **b**, the ion size

distribution as seen by the NAIS ion mode; and **c**, the monomer (light green) and dimer (dark green) and the number particle concentration for particles bigger than 3.2 nm (purple; CPC 3776). The colour keys on the right side relate to the number size distribution (dN/dlogD_p).



Extended Data Figure 9 | Annually averaged HOM concentration, and the annually averaged growth rate, simulated by GLOMAP at cloud base level. a, Spatial distribution of HOM concentration (in cm^{-3}). **b–d**, Spatial distribution of growth rates (in nm h^{-1}) using different parameterizations: **b**, using the size-dependent parameterization of initial

particle growth and irreversible condensation of H_2SO_4 , **c**, with growth from 1.7 nm to 3 nm only due to H_2SO_4 , and **d**, with growth from 1.7 nm to 3 nm assuming irreversible condensation of H_2SO_4 together with an organic contribution following ref. 30, which assumes a Kelvin barrier to organic condensation below 2.5 nm.

Extended Data Table 1 | Summary of CLOUD runs during CLOUD7 and CLOUD8

CLOUD 7					
RUN	Ozone [p.p.b.v.]	α -pinene [p.p.t.v.]	SO ₂ [p.p.b.v.] [‡]	HOM [cm ⁻³]	Sulfuric Acid [cm ⁻³] [‡]
1060	20	660	70	$1.1 \cdot 10^7$	$1.9 \cdot 10^7$
1061	20	640	70	$1.0 \cdot 10^7$	$1.8 \cdot 10^7$
1062	20	180	70	$4.2 \cdot 10^6$	$7.8 \cdot 10^6$
1063	20	650	70	$1.0 \cdot 10^7$	$1.6 \cdot 10^7$
1065	20	190	70	$3.8 \cdot 10^6$	$6.5 \cdot 10^6$
1066	20	650	70	$1.3 \cdot 10^7$	$8.7 \cdot 10^6$
1067	20	640	70	$1.3 \cdot 10^7$	$6.6 \cdot 10^6$
1068	20	890	70	$1.5 \cdot 10^7$	$1.2 \cdot 10^7$
1070	20	1230	70	$1.7 \cdot 10^7$	$2.4 \cdot 10^7$
1107	30	420	0.6	$1.4 \cdot 10^7$	$7.6 \cdot 10^5$
1108	30	420	0.6	$1.4 \cdot 10^7$	$6.0 \cdot 10^5$
1109	30	430	0.6	$1.4 \cdot 10^7$	$6.3 \cdot 10^6$
1110	30	430	0.6	$1.3 \cdot 10^7$	$5.1 \cdot 10^6$
1111	30	430	1.6	$1.4 \cdot 10^7$	$5.6 \cdot 10^6$
1113	30	370	1.6	$1.3 \cdot 10^7$	$5.5 \cdot 10^6$
1114	30	390	1.6	$5.8 \cdot 10^6$	$4.3 \cdot 10^6$

[‡] measured with SO₂ monitor (Thermo scientific, 43i-TLE)[†] measured with CIMS

CLOUD 8					
RUN	Ozone [p.p.b.v.]	α -pinene [p.p.t.v.]	SO ₂ [p.p.b.v.]	HOM [cm ⁻³]	Sulfuric Acid [cm ⁻³]
1208	32	20	–°	$2.9 \cdot 10^6$	$6.8 \cdot 10^4$
	34	40	–°	$5.8 \cdot 10^6$	$6 \cdot 10^4$
	34	110	–°	$1.4 \cdot 10^7$	$7.4 \cdot 10^4$
1209	35	250	–°	$2.0 \cdot 10^7$	$3.8 \cdot 10^4$
1210	35	510	–°	$3.4 \cdot 10^7$	$3.4 \cdot 10^4$
1211	34	240	5	$1.7 \cdot 10^7$	–*
1212	32	1340	20	$1.0 \cdot 10^8$	$7.5 \cdot 10^5$
1213	33	1280	10	$7.4 \cdot 10^7$	$1.3 \cdot 10^5$
1214	33	170	8	$1.4 \cdot 10^7$	$8.2 \cdot 10^4$
1215	33	90	6	$7.4 \cdot 10^6$	$6.3 \cdot 10^4$
1217	32	270	40	$1.9 \cdot 10^7$	$1.1 \cdot 10^5$
1218	33	250	150	$1.7 \cdot 10^7$	$4.4 \cdot 10^5$
1219	34	250	1300	$1.4 \cdot 10^7$	$2.9 \cdot 10^6$
1220	33	240	1500	$1.2 \cdot 10^7$	$3.3 \cdot 10^6$
1221	31	11000	20	$2.1 \cdot 10^8$	$3.1 \cdot 10^4$
1222	31	7750	20	$1.8 \cdot 10^8$	$2.9 \cdot 10^4$
1224	33	900	–*	$4.2 \cdot 10^7$	$9.4 \cdot 10^5$
1225	31	15700	–*	$2.5 \cdot 10^8$	$2.6 \cdot 10^5$
1226	31	60	–*	$3.8 \cdot 10^6$	$8.2 \cdot 10^4$
1226	30	700	–*	$2.8 \cdot 10^7$	$5.1 \cdot 10^4$
1227	33	460	–*	$1.9 \cdot 10^7$	$3.1 \cdot 10^4$
1229	33	290	4000	$9.2 \cdot 10^6$	$3.1 \cdot 10^7$

* no measurement available

° no SO₂ injected

Each run consisted of several stages (increasing gases, steady-state, changing charging state of chamber, see also Extended Data Fig. 8), here only the steady-state plateau values are indicated.

Competitive growth in a cooperative mammal

Elise Huchard^{1,2}, Sinead English^{1†}, Matt B. V. Bell^{1†}, Nathan Thavarajah³ & Tim Clutton-Brock^{1,3}

In many animal societies where hierarchies govern access to reproduction, the social rank of individuals is related to their age and weight^{1–5} and slow-growing animals may lose their place in breeding queues to younger ‘challengers’ that grow faster^{5,6}. The threat of being displaced might be expected to favour the evolution of competitive growth strategies, where individuals increase their own rate of growth in response to increases in the growth of potential rivals. Although growth rates have been shown to vary in relation to changes in the social environment in several vertebrates including fish^{2,3,7} and mammals⁸, it is not yet known whether individuals increase their growth rates in response to increases in the growth of particular reproductive rivals. Here we show that, in wild Kalahari meerkats (*Suricata suricatta*), subordinates of both sexes respond to experimentally induced increases in the growth of same-sex rivals by raising their own growth rate and food intake. In addition, when individuals acquire dominant status, they show a secondary period of accelerated growth whose magnitude increases if the difference between their own weight and that of the heaviest subordinate of the same sex in their group is small. Our results show that individuals adjust their growth to the size of their closest competitor and raise the possibility that similar plastic responses to the risk of competition may occur in other social mammals, including domestic animals and primates.

Recent studies have revealed the extent to which aspects of the social environment can affect growth in several vertebrates. In some social fish, the risk of conflict with dominant individuals reduces the growth rates of subordinates^{2,3,7} while, in some mammals, prenatal growth increases in response to physiological stress levels in pregnant mothers in high-density environments⁸. However, studies have not yet investigated whether adolescents or adults can adjust their growth rates in relation to changes in the size of specific rivals that may displace them in reproductive queues. In many cooperatively breeding mammals, subordinates of both sexes queue for reproductive opportunities in breeding groups, sometimes for several years^{5,9}. Rank in these queues is usually determined by relative age and weight, and previous research has produced some evidence of strategic adjustments in growth. In mole-rats and meerkats, adult females that acquire the dominant breeding position commonly show a period of secondary growth^{10–12} which may allow them to increase their fertility or consolidate their status^{5,13}. Here, we describe experiments that investigate whether subordinate meerkats queuing for breeding opportunities also engage in competitive growth.

Meerkats live in groups of 3–50 individuals where 90% of reproduction is monopolized by a single dominant pair⁵. Subordinates of both sexes contribute to costly cooperative activities, including pup-feeding, babysitting and raised-guarding¹⁴. Within groups, subordinates of the same sex are ranked in a hierarchy based on age and weight¹⁵. If the breeding female dies, the oldest and heaviest subordinate typically replaces her, and subordinate females occasionally displace breeders⁵. Unlike females, most males leave their natal groups voluntarily when they are 2–4 years old in small parties of two to six individuals, and attempt to displace males in other groups^{5,16}. If they are successful,

the oldest and heaviest male in the party often assumes the breeding position. Data presented here are derived from a 24-year study of wild meerkats that has encompassed more than 60 groups in which all individuals were recognizable. Most individuals were trained to climb onto electronic balances and were weighed three times a day (dawn; after 3 h of foraging; and dusk) on approximately 10 days a month throughout their lives⁵. Changes in the weight of individuals between the beginning and end of morning foraging sessions provide a measure of their food intake.

Using 14 groups of habituated meerkats, we manipulated the growth of subordinates of both sexes by provisioning particular individuals and measuring effects on the growth and food intake of individuals of the same sex immediately above them in the age-related hierarchy. We identified pairs of same-sex littermates belonging to two distinct age classes: juveniles (aged 4–7 months), which had recently reached nutritional independence ($n = 12$ female and 19 male litters from 12 groups), and young adults (aged 12–24 months), which had reached sexual maturity and were able to compete for any breeding vacancies that occurred⁵ ($n = 8$ female and 9 male litters from 14 groups). In each pair, we fed the lighter individual, later referred to as the ‘challenger’, with half a hard-boiled egg twice per day for 3 months. We subsequently compared the growth of unfed littermates, referred to as ‘challenged’ individuals, with those of unfed control individuals of the same age from other litters over the same period (Extended Data Fig. 1).

Challenged individuals of both age classes responded to increases in the growth of fed challengers by increasing their average weight (both in absolute terms and relative to controls) over the course of the experiment. Growth from the start to the mid-point of the experiment was greater in challenged than in control individuals (Fig. 1a, b; juveniles: two sample Welch’s t -test, $n = 32$ challenged and 72 control individuals, $t = 4.17$, $P < 10^{-4}$; adults: $n = 18$ challenged and 18 age- and sex-matched control individuals, paired t -test, $t = 2.10$, d.f. = 17, $P = 0.050$), generating a difference in the average weight of challenged and control individuals halfway through the experiment (juveniles: $n = 32$ challenged and 83 control individuals, 504.3 ± 68.2 g versus 438.5 ± 73.2 g, two-sample Welch’s t -test, $t = 4.54$, $P < 10^{-4}$; adults: pairwise weight difference = 40.7 ± 51.06 g, paired t -test, $t = 3.38$, d.f. = 17, $P = 0.003$). Differences in growth were, however, no longer detectable in the second half of the experiment (juveniles: $n = 27$ challenged and 74 control individuals, two-sample Welch’s t -test, $t = 0.22$, $P = 0.825$; adults: paired t -test, $t = -24.23$, d.f. = 17, $P = 0.059$), suggesting that challenged individuals may not be capable of sustaining accelerated growth over extended periods. In both age classes, the growth of challenged individuals over the first half of the experiment was positively correlated with the growth of their fed challenger (Extended Data Fig. 2 and Extended Data Table 1), suggesting that challenged individuals adjusted their growth response to the growth of their rival. Increases in the growth of challenged individuals were associated with increases in food intake: food intake was greater for challenged than for control individuals in the first half of the experiment (Fig. 1c, d; juveniles: $n = 32$ challenged and 86 control individuals, two-sample Welch’s t -test, $t = 2.17$, $P = 0.033$; adults: paired t -test: $t = 2.80$, d.f. = 16, $P = 0.013$),

¹Large Animal Research Group, Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK. ²CEFE UMR 5175, CNRS - Université de Montpellier, 1919 Route de Mende, 34293 Montpellier Cedex 5, France. ³Department of Zoology and Entomology, Mammal Research Institute, University of Pretoria, Pretoria, Gauteng 0002, South Africa. [†]Present addresses: Behavioural Ecology Group, Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK (S.E.); Institute for Evolutionary Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK (M.B.V.B.).

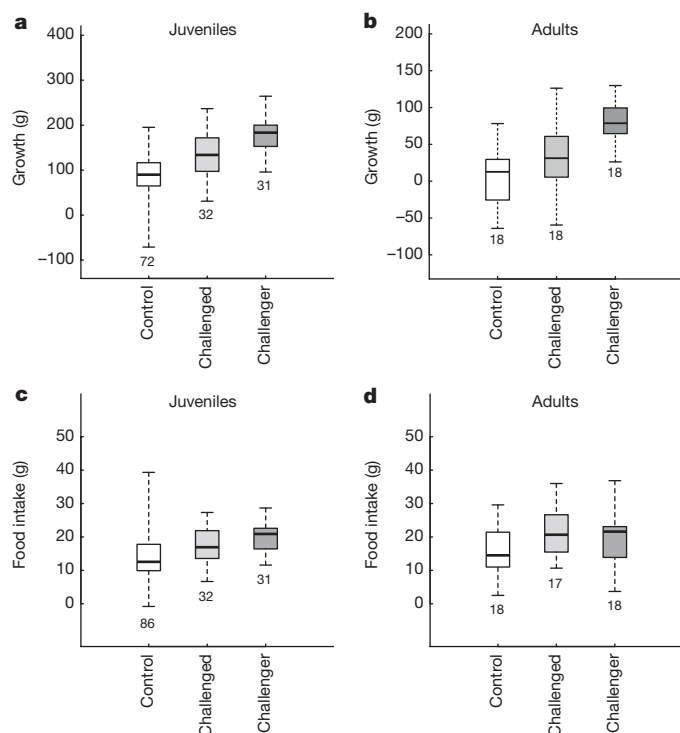


Figure 1 | Competitive growth in subordinates. Boxplots showing the growth (individual weight difference between the start and mid-point of the experiment) (a, b) and food intake (average morning weight gain in the first half of experiment) (c, d) of unfed, 'challenged' individuals (light grey boxes) and of their fed 'challengers' (dark grey boxes) relative to control individuals (white boxes) in juveniles (a, c) and adults (b, d). Whiskers comprise all data points. Numbers below the boxes indicate the number of individuals.

but not in the second half (juveniles: $n = 29$ challenged and 83 control individuals, two-sample Welch's t -test, $t = 1.19$, $P = 0.240$; adults: paired t -test: $t = -0.16$, d.f. = 16, $P = 0.876$).

Social mechanisms other than competitive growth could conceivably contribute to increases in the growth of challenged animals, but we were unable to find any evidence that this was the case. It is unlikely that potential increases in the contributions of fed challengers to cooperative activities in the first half of experiment reduced the contributions of challenged animals and so increased their weight gain. First, juveniles contribute little to cooperative activities, so accelerated growth in challenged juveniles cannot be mediated by changes in cooperative behaviour. Second, challenged adults maintained their investment in raised-guarding and pup-feeding in the same period relative to control animals (Wilcoxon signed-rank paired-test; raised-guarding: $V = 52$, d.f. = 17, $P = 0.156$; pup-feeding: $V = 30$, d.f. = 14, $P = 0.095$). Finally, adult fed challengers increased their contributions to raised-guarding but not to pup-feeding (Wilcoxon signed-rank paired-test; raised-guarding: $V = 143$, d.f. = 17, $P = 0.013$; pup-feeding: $V = 67$, d.f. = 14, $P = 0.719$).

Additional analyses suggest that adults that acquire dominant positions may also adjust their growth rates in a strategic fashion. In both sexes, the lifetime breeding success of dominant meerkats depends on the length of time they hold the dominant position⁵ which, in females, increases with the difference between their own weight and the weight of the heaviest subordinate of the same sex⁵. Since subordinates engage in competitive growth, we examined whether individuals that have recently acquired the dominant position adjust the magnitude of their subsequent increase in weight to the relative weight of their closest rival. We first analysed whether newly dominant males and females increase their growth rate following dominance acquisition by comparing their weight in the month before dominance acquisition and

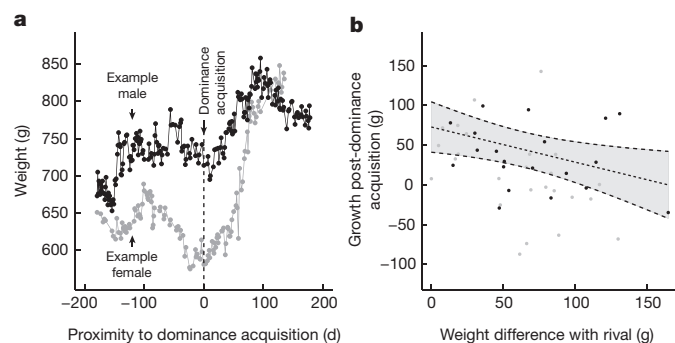


Figure 2 | Competitive growth in dominants. a, Example growth trajectories of a male and female during their transition to dominance. b, Adjustment of growth following dominance acquisition in response to social competition in 20 males and 25 females. Dots show the raw values (grey for females, black for males) of dominant weight gain within the 150 days following dominance acquisition as a function of weight difference to the heaviest same-sex subordinate (measured at dominance acquisition). The dotted line shows the predicted values of the linear model (results presented in Extended Data Table 3); the s.d. of the predicted values are delineated by shaded areas.

in the 4 months following dominance acquisition. New dominants of both sexes increased in weight after acquiring dominance (analysis of variance with repeated measures; effect of month post-dominance acquisition on weight: $F_{4,184} = 16.81$, $P < 10^{-4}$; Fig. 2a and Extended Data Fig. 3a). The extent of growth following dominance acquisition did not differ between the sexes (analysis of variance with repeated measures; interaction between sex and month post-dominance acquisition: $F_{4,184} = 1.22$, $P = 0.31$) and occurred primarily in the 2 months following dominance acquisition (see Extended Data Table 2 for the results of the post-hoc tests). This growth response may not solely reflect improved access to resources, as food intake remained constant in both sexes during the same period (analysis of variance with repeated measures; effect of month post-dominance acquisition on food intake: $F_{4,112} = 0.34$, $P = 0.850$; interaction between sex and month post-dominance acquisition: $F_{4,112} = 0.09$, $P = 0.986$; Extended Data Fig. 3b).

The growth of new dominants in the 5 months following dominance acquisition was more pronounced when the heaviest same-sex subordinate was closer to their own weight at the time of dominance acquisition (linear model; estimate \pm s.d. = -0.76 ± 0.27 , $F_{1,36} = 7.69$, $P < 0.01$; Fig. 2b and Extended Data Table 3). There was no significant sex difference in this accelerated growth (Extended Data Table 3). Rapid post-dominance growth exacerbated existing weight differences between dominants and same-sex subordinates, with the result that most established dominants were the heaviest individual of their sex in their group (females: 58% of groups; males: 68%). While similar periods of growth after dominance acquisition in female naked mole-rats have been interpreted as a way of enhancing fecundity^{11,12,17}, the presence of strategic growth adjustments to the relative size of rivals in dominant meerkats of both sexes suggests that these increases may serve to consolidate their status and prolong their breeding tenure^{5,13}.

Our findings suggest that subordinates can track changes in the growth and size of potential competitors, perhaps using physical contact as well as visual, vocal or olfactory cues, and react by adjusting their own growth. While the physiological correlates of increased growth rates in challenged individuals are not yet known, hormonal changes associated with heightened threat of competition may increase growth and food intake. Acceleration in growth following dominance acquisition is probably associated with the sudden lifting of reproductive suppression and a re-orientation of life-history strategy. The hormonal profile of dominant meerkats is distinct from that of subordinates, with higher plasmatic levels of oestradiol and progesterone in breeding females and of cortisol in breeders of both sexes^{10,18,19}. Sex steroids

are known to regulate the production of critical actors in the insulin/growth factor pathway in the mammalian reproductive tract and associated tissues²⁰, which may result in the upregulation of anabolic genes involved in growth. Strategic increases in growth rates could be constrained by energy and fitness costs²¹. Allocation of additional resources to growth by challenged individuals may depress immune function and reduce longevity as a result of increases in oxidative stress and telomere shortening²² while increases in time spent foraging may raise predation risk, which is high in meerkats²³.

Our results suggest that competitive growth may represent an important component of the developmental strategy of individuals. Recognition of this process may alter classic perspectives on mechanisms of social competition, which frequently suggest that the phenotype of interacting individuals determines the outcome of competitive interactions rather than vice versa. As reproductive queues are widespread in social mammals and the size and weight of individuals often affect their status and breeding success²⁴, competitive growth may occur in many other social species, possibly including domestic mammals, non-human primates and humans.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 July 2015; accepted 6 April 2016.

- Hoogland, J. L. *The Black-Tailed Prairie Dog: Social Life of a Burrowing Mammal* (Univ. Chicago Press, 1995).
- Buston, P. Social hierarchies: size and growth modification in clownfish. *Nature* **424**, 145–146 (2003).
- Heg, D., Bender, N. & Hamilton, I. Strategic growth decisions in helper cichlids. *Proc. R. Soc. Lond. B* **271**, S505–S508 (2004).
- Spong, G. F., Hodge, S. J., Young, A. J. & Clutton-Brock, T. H. Factors affecting the reproductive success of dominant male meerkats. *Mol. Ecol.* **17**, 2287–2299 (2008).
- Clutton-Brock, T. H. et al. Intrasexual competition and sexual selection in cooperative mammals. *Nature* **444**, 1065–1068 (2006).
- Reeve, H. K., Peters, J. M., Nonacs, P. & Starks, P. T. Dispersal of first “workers” in social wasps: causes and implications of an alternative reproductive strategy. *Proc. Natl Acad. Sci. USA* **95**, 13737–13742 (1998).
- Wong, M. Y. L., Munday, P. L., Buston, P. M. & Jones, G. P. Fasting or feasting in a fish social hierarchy. *Curr. Biol.* **18**, R372–R373 (2008).
- Dantzer, B. et al. Density triggers maternal hormones that increase adaptive offspring growth in a wild mammal. *Science* **340**, 1215–1217 (2013).
- Hauber, M. E. & Lacey, E. A. Bateman’s principle in cooperatively breeding vertebrates: the effects of non-breeding alloparents on variability in female and male reproductive success. *Integr. Comp. Biol.* **45**, 903–914 (2005).
- Russell, A. F., Carlson, A. A., McIlrath, G. M., Jordan, N. R. & Clutton-Brock, T. Adaptive size modification by dominant female meerkats. *Evolution* **58**, 1600–1607 (2004).
- Young, A. J. & Bennett, N. C. Morphological divergence of breeders and helpers in wild Damaraland mole-rat societies. *Evolution* **64**, 3190–3197 (2010).
- Dengler-Criss, C. M. & Catania, K. C. Phenotypic plasticity in female naked mole-rats after removal from reproductive suppression. *J. Exp. Biol.* **210**, 4351–4358 (2007).
- Clutton-Brock, T. Structure and function in mammalian societies. *Phil. Trans. R. Soc. B* **364**, 3229–3242 (2009).
- Clutton-Brock, T. H. et al. Evolution and development of sex differences in cooperative behavior in meerkats. *Science* **297**, 253–256 (2002).
- Thavarajah, N. K., Fenkes, M. & Clutton-Brock, T. H. The determinants of dominance relationships among subordinate females in the cooperatively breeding meerkat. *Behaviour* **151**, 89–102 (2014).
- Doolan, S. P. & Macdonald, D. W. Dispersal and extra-territorial prospecting by slender-tailed meerkats (*Suricata suricatta*) in the south-western Kalahari. *J. Zool.* **240**, 59–73 (1996).
- O’Riain, M. J., Jarvis, J. U., Alexander, R., Buffenstein, R. & Peeters, C. Morphological castes in a vertebrate. *Proc. Natl Acad. Sci. USA* **97**, 13194–13197 (2000).
- Carlson, A. A. et al. Hormonal correlates of dominance in meerkats (*Suricata suricatta*). *Horm. Behav.* **46**, 141–150 (2004).
- Young, A. J., Monfort, S. L. & Clutton-Brock, T. H. The causes of physiological suppression among female meerkats: a role for subordinate restraint due to the threat of infanticide? *Horm. Behav.* **53**, 131–139 (2008).
- Dantzer, B. & Swanson, E. M. Mediation of vertebrate life histories via insulin-like growth factor-1. *Biol. Rev. Camb. Phil. Soc.* **87**, 414–429 (2012).
- Arendt, J. D. Adaptive intrinsic growth rates: an integration across taxa. *Q. Rev. Biol.* **72**, 149–177 (1997).
- Metcalfe, N. B. & Monaghan, P. Compensation for a bad start: grow now, pay later? *Trends Ecol. Evol.* **16**, 254–260 (2001).
- Clutton-Brock, T. H. et al. Predation, group size and mortality in a cooperative mongoose, *Suricata suricatta*. *J. Anim. Ecol.* **68**, 672–683 (1999).
- Clutton-Brock, T. H. & Huchard, E. Social competition and selection in males and females. *Phil. Trans. R. Soc. B* **368**, 20130074 (2013).

Acknowledgements We are grateful to the many volunteers, field managers, PhD students and post-doctoral researchers who have contributed to data collection over the past 15 years, and to D. Gaynor, I. Stevenson, P. Roth, J. Samson, R. Millar, E. Cameron, J. du Toit and M. Haupt for support. We are grateful to M. Manser for her contribution to the organization of the Kalahari Meerkat Project and to C. Drea for additional help and advice. We also thank D. Cram for comments on previous drafts, and A. Bateman, A. Courtiol and M. Crawley for statistical advice. Northern Cape Conservation and the Kotze family provided permission to work in the Kalahari. Our work was approved by the Animal Ethics Committee of the University of Pretoria (project number EC010-13). The Kalahari Meerkat Project is supported and organized by the Universities of Cambridge and Zurich. This research was supported by the Natural Environment Research Council (grant NE/G006822/1) and the European Research Council (grant 294494).

Author Contributions E.H. implemented the analysis and drafted the results; T.H.C.-B., S.E. and M.B. planned the experiments, which were conducted by N.T. and other members of the Kalahari Meerkat Project; E.H., S.E., M.B. and T.H.C.-B. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.H. (ehuchard@gmail.com or elise.huchard@cefe.cnrs.fr).

METHODS

Study site and population. Data were collected between 1996 and 2013 as part of a long-term study of wild meerkats at the Kuruman River Reserve, South Africa. The site experiences a hot–wet season (October–April) and a cold–dry season (May–September), with extensive inter-annual variation in rain²³. Rainfall was measured daily (in millimetres) using a standard gauge²⁵. Details about the site and population are published elsewhere^{5,14,23}.

Meerkats were habituated to humans and individually recognizable by dye marks. Groups were visited about three times a week, so life-history events (births, deaths, emigrations, changes in dominance) were known to an accuracy of about 3 days (refs 5, 14). Pregnancy status was inferred from parturition date and affects female weight from the midpoint of gestation, lasting approximately 70 days (ref. 26). Females were considered pregnant from 40 days before parturition or from the first day of detectable pregnancy in cases where abortions occurred. Dominant individuals were identified by their behaviour towards group-mates^{4,5}. They scent-marked more frequently than subordinates, and asserted their dominance over others by anal marking, by rubbing them with their chin, and more rarely by attacking and biting them. Changes in dominance were immediately recognizable, as they were often preceded by a short period (hours to days) of intense fighting, and were accompanied by dramatic changes in behaviour in the contesting individuals. Previous genetic work has shown the absence of incestuous matings within groups⁴. If all immigrant males die, a natal male may become socially dominant in his group. Natal dominant males do not mate-guard the dominant female, which is often their mother, and regularly conduct extraterritorial forays for mating opportunities²⁷. These males (77/166 dominant males in our dataset) were excluded from analyses.

Weight measures. Individuals were trained to climb onto a laboratory balance in return for drops of water or crumbs of hard-boiled egg, allowing us to record body weight to an accuracy of 1 g. Although individuals were often weighed three times a day, we only used data collected in the morning right after emergence from the burrow and before foraging, to avoid noise created by variation in foraging success throughout the day²⁵. Food intake, or morning weight gain, was calculated as the difference between weight collected before foraging activity started, and weight collected after about 3 h of foraging¹⁰.

Cooperative behaviour. Three cooperative activities are regularly performed by male and female meerkats¹⁴: (1) babysitting newborn pups, where an individual stays at the burrow while the rest of the group forages; (2) feeding pups that are old enough to join foraging trips (approximately 1–3 months old); and (3) raised-guarding, where an individual ceases foraging and climbs to a raised position to watch out for potential dangers. The occurrence of babysitting, pup-feeding and raised-guarding was recorded *ad libitum* as events during observation sessions, allowing quantification of relative rates of helping per individual: that is, the number of occurrences of one cooperative behaviour performed by one individual relative to the total number of occurrences of that behaviour in the group over a given period.

Competitive growth experiment. From 2010 to 2013, we conducted a set of 3-month feeding experiments on adults aged 310–870 days and on juveniles aged 111–215 days to investigate whether unfed littermates (challenged individuals) would increase their growth rate in response to experimentally elevated growth rates of their fed siblings (challengers). We identified pairs containing at least two same-sex littermates and fed the individual that was lightest (or as heavy as its sibling) when the experiment started (mean weight difference (\pm s.d.) in juveniles: 9.8 ± 30.6 g; in adults: 29.9 ± 28.2 g). The fed individuals received half an egg twice daily four times a week for 3 months. Competitive growth has never been described previously, so no prior information was available for power analyses to establish adequate sample sizes. For 17 fed adults including 8 females, the shortest feeding bout lasted 55 days and the mean \pm s.d. feeding duration was 84 ± 11 days. For 31 fed juveniles including 12 females, the shortest feeding bout lasted 21 days and the mean \pm s.d. feeding duration was 76 ± 21 days. For one adult female litter and one juvenile male litter, there were three same-sex siblings and the two lightest individuals were very close in weight (that is, their average weight difference was lower than 10 g in the 15 days preceding the experiment); one of them was fed, and the two unfed siblings were included in the cohort of challenged individuals. Experiments were interrupted when a pregnancy was detected in an experimental female (fed or unfed), and corresponding data were excluded from analysis. In other cases where the experiment was aborted (for example, if an individual disappeared), data collected during the shortened period were included in analyses; note that for three juvenile dyads, food supplementation lasted respectively 21, 23 and 26 days, so these individuals were excluded from all calculations related to measures describing the second half of the experiment. Observations and weighing sessions were not subjected to blinding, because weight gained by fed individuals during the experiment was often detectable by observers.

Statistical analysis. To investigate the effect of feeding individuals on the growth of their unfed same-sex littermate, we first calculated the growth and food

intake, averaged over the first or the second half of the experiment for challenged individuals, challengers and control individuals. Growth was calculated as the individual difference between weight recorded immediately before the start of the experiment and at the mid-point of the experiment (45 days), or as the individual change in weight from the mid-point to the end of the experiment (90 days). Food intake, calculated in terms of morning weight gain, was averaged for each individual, over days 5–45 of the experiment (the first 4 days were excluded to allow for potential adjustments in challenged individuals) and then over experimental days 45–90. We compared these measures across challenged and control individuals using two-sample Welch's *t*-tests (for juveniles) and paired *t*-tests (for adults) after checking that variance was homogeneous across groups using Levene tests ($P > 0.05$ in all cases). We focused on the contrast between challenged and control individuals: significantly higher growth in challenged individuals over controls would provide experimental evidence for competitive growth, defined as an elevated increase in growth in response to the challenge of a fed rival. Control individuals were selected as any individual from the population during the experimental period (2010–2013) that had a lighter same-sex littermate in their group at the age at which supplemental feeding started in experimental groups (120 days in juveniles, 1 year in adults), to match criteria used to identify unfed individuals in experimental dyads (Extended Data Fig. 1). In adults, where heterogeneity in the age at the start of the experiment was considerable ($361\text{--}772$ days, mean \pm s.d. = 496.7 ± 112.9 days), each challenged individual was matched to the same-sex individual of the control cohort that was closest in age (differences in birth dates between challenged individuals and their matched control were small: $2\text{--}32$ days, mean \pm s.d. = 11.2 ± 8.4) and present in the population at the time of the experiment. Matching each experimental individual with a same-age and same-sex control in this way allowed us to control for environmental variation that might otherwise have introduced noise when comparing the weight and growth of individuals that underwent a supplementation at different periods (e.g. during the dry versus the wet season). Individual weight before the experiment was averaged across the 15 days preceding the experiment; weight at mid-point was averaged across days 45–60 of the experiment; and weight at the end of the experiment was averaged across experimental days 90–105.

It was not possible to select such matched control individuals in juveniles, however, as there was no control litter born shortly before or after experimental litters in several cases. Small age differences can introduce important noise when comparing weights among juveniles, because growth rates are relatively high between 4 and 7 months of age, compared with later ages²⁵. In the juvenile cohort, age at the start of the experiment was very homogeneous (range: $111\text{--}128$ days of age, mean \pm s.d. = 122.3 ± 4.7), so matching experimental dyads with control individuals by age was deemed less necessary. Individual weight records were averaged across 95–110 days of age (before experiment); 170–185 days of age (after about 45 days of experiment); and 215–230 days of age (after about 90 days of experiment), and growth was calculated between these time points.

We further ran a linear model investigating the relationship between the growth of challenged individuals and the growth of their fed challenger to test whether the growth responses of challenged individuals were adjusted to the weight gain of their fed challenger. Growth was the response variable, and was calculated as the weight difference between the start and the mid-point of the experiment (since the above analyses suggested that competitive growth was highest at this time). Explanatory variables included sex, age at start of experiment and cumulative rainfall in the previous 9 months, which was previously found to influence the growth of individual meerkats²⁵. Results and sample sizes are presented in Extended Data Table 1 and Extended Data Fig. 2.

We investigated the influence of the experiment on pup-feeding and raised-guarding rates in the adult cohort only, because helping is rare before 6 months of age¹⁴. We did not consider babysitting because fewer than half of the experimental groups exhibited babysitting during the experiment. For each observation session, we measured the observed proportion of raised-guarding events performed by the focal individual relative to the total number of events recorded for the group. We then calculated individual deviation from the proportion expected under the null hypothesis, where each individual contributes equally, calculated as the inverse of the number of helpers in the group. We averaged this deviation across all observation sessions for each individual during the first half of the experiment (10–120 sessions per individual, median = 19). Thus, mean deviation gives an indication of the extent of cooperative behaviour relative to average contributions in the group: individuals with a larger, more positive deviation have higher cooperative behaviour. We compared the mean deviations between challenged individuals and their matched controls using paired Wilcoxon signed-rank tests, as the response variable was not normally distributed. We used the same approach to test for differences in individual contributions to pup-feeding between challenged and control individuals.

When investigating changes in weight following dominance acquisition, we considered individuals that maintained dominance for at least 6 months, to avoid biasing the sample towards short and unstable tenures. We averaged weight records for each individual ($n = 42$ females and 30 males) across the 30 days preceding dominance acquisition (labelled 'month 0') and then across days 0–30, 30–60, 60–90 and 90–120 following dominance acquisition (respectively labelled 'months 1, 2, 3 and 4'). Weights recorded during pregnancies were excluded. We then retained only individuals with no missing data in any of these five 1-month blocks ($n = 21$ females and 27 males) to ensure a balanced design. Thus, we could evaluate the significance of weight differences between 1-month blocks using a repeated-measures analysis of variance with multiple factors. Factors included sex, proximity to dominance acquisition (with five levels: month 0, 1, 2, 3 and 4) and the interaction between sex and proximity to dominance acquisition, to test if the temporal dynamics of post-dominance growth differed between males and females. Post-hoc tests were conducted using paired t -tests with adjusted P values to compare within-individual changes in weight before dominance acquisition to each of the 4 months after acquisition; as well as between each month of the 4-month period following acquisition of dominance. A Bonferroni correction was applied to correct for multiple testing. These results are presented in Extended Data Fig. 3a and Extended Data Table 2.

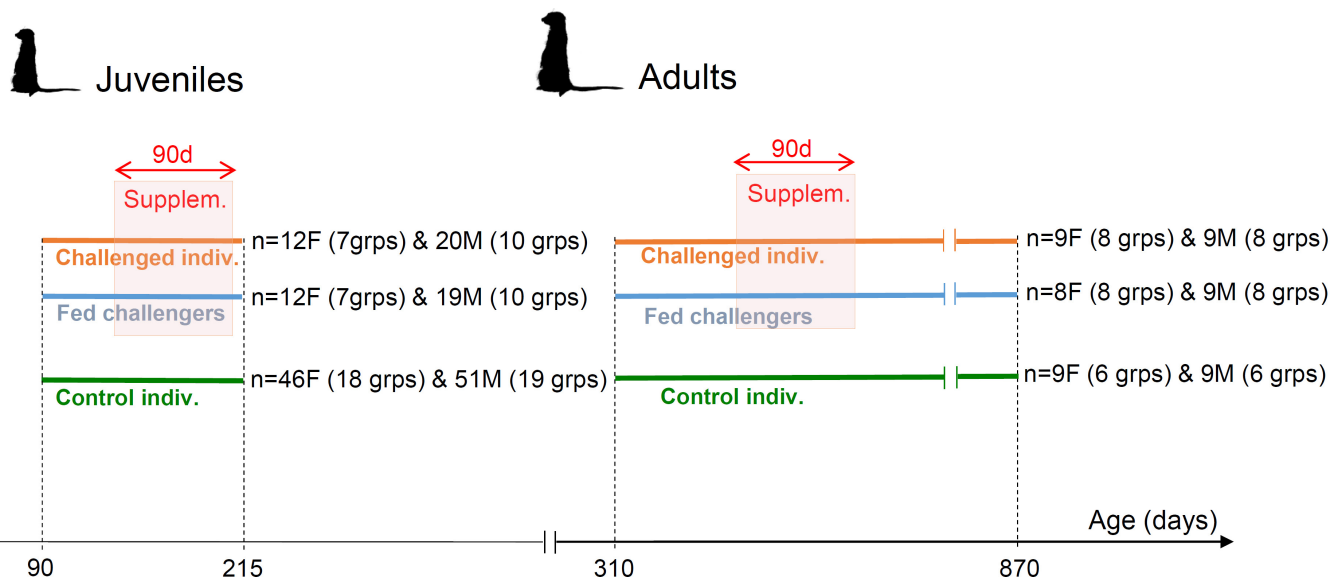
We compared changes in food intake (measured as morning weight gain) following dominance acquisition using the same approach. As described above, we retained only individuals with no missing data in any of the five 1-month blocks ($n = 9$ females and 21 males) to evaluate the significance of differences in food intake between 1-month blocks using a repeated-measures analysis of variance with multiple factors. As above, factors included were sex, proximity to dominance acquisition and their interaction. These results are illustrated in Extended Data Fig. 3b.

To investigate the effect of competition on growth following dominance acquisition, we ran a linear model, with weight gain within 150 days following dominance acquisition (calculated as weight 150 days after dominance acquisition minus weight at dominance acquisition, each averaged across all weights for 10 days before and after the time-point of interest) as our response variable. We focused on a 5-month period after dominance acquisition, because previous analyses had

revealed that growth rates were elevated in the 2–4 months following dominance acquisition. We included all new dominant females that retained dominance for longer than 6 months and had at least one subordinate female in their group that was older than 6 months when they became dominant. Six months is the age of the youngest female that ever reached dominance. Weights recorded during pregnancies were excluded. We included all new dominant males that had at least one non-natal subordinate male in their group that was older than 6 months when they became dominant. Natal subordinate males were not considered as rivals because they hardly ever reproduce or fight for dominance⁴. Explanatory variables included sex, rainfall (averaged over the 150 days following dominance acquisition), a sinusoidal term describing season of dominance acquisition²⁵, age at dominance acquisition, and absolute weight difference with the same-sex rival (that is, heaviest subordinate at the time of dominance acquisition). In addition, the interaction between sex and absolute weight difference with the same-sex rival tested whether the effect of the weight difference with the main rival differed between sexes. We used the absolute value of weight difference because graphical exploration of the data suggested that dominant growth rates increase when the main same-sex rival is either slightly heavier or slightly lighter, but not when the rival is much lighter or much heavier. In cases where a rival is much heavier but fails to win fights over dominance, he or she may have poor competitive abilities for other reasons and may not represent a threat to the dominant. The results and sample sizes are presented in Extended Data Table 3.

All statistical analyses were run with R 3.1.3 (ref. 28), and all tests were two-sided.

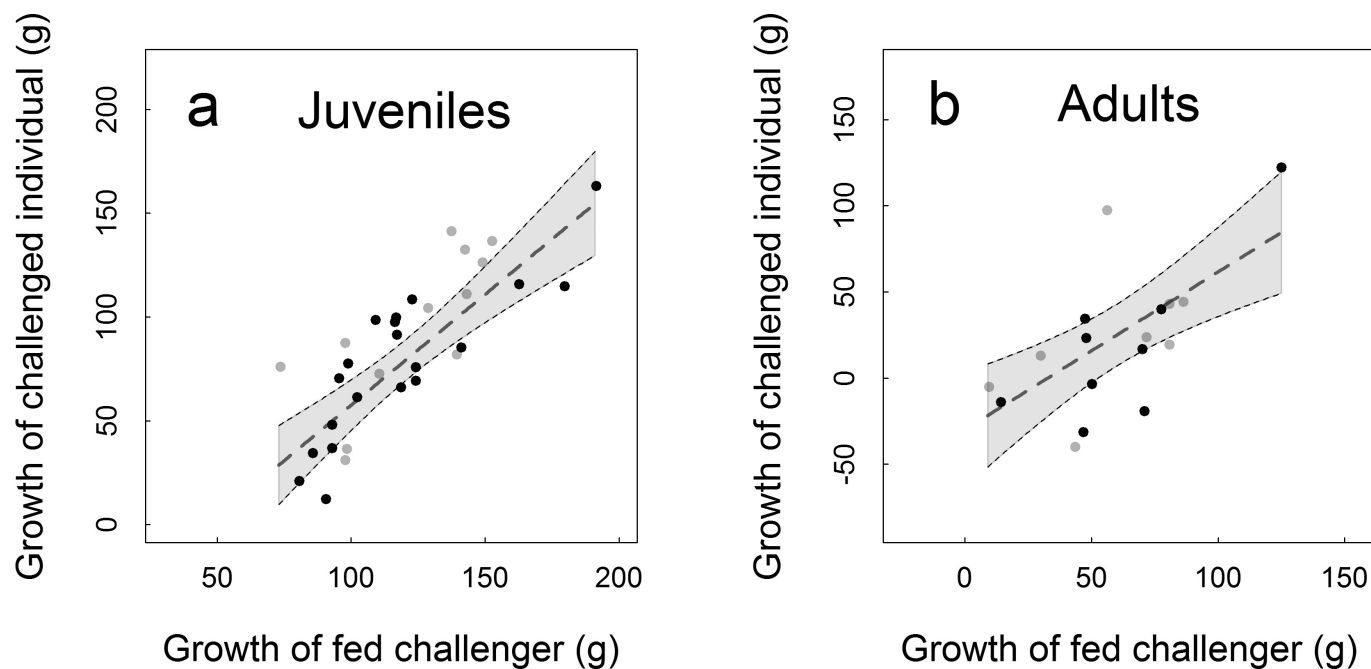
25. English, S., Bateman, A. W. & Clutton-Brock, T. H. Lifetime growth in wild meerkats: incorporating life history and environmental factors into a standard growth model. *Oecologia* **169**, 143–153 (2012).
26. Sharp, S. P., English, S. & Clutton-Brock, T. H. Maternal investment during pregnancy in wild meerkats. *Evol. Ecol.* **27**, 1033 (2013).
27. Young, A. J., Spong, G. & Clutton-Brock, T. Subordinate male meerkats prospect for extra-group paternity: alternative reproductive tactics in a cooperative mammal. *Proc. R. Soc. B* **274**, 1603–1609 (2007).
28. R Development Core Team. R: A Language and Environment for Statistical Computing. <http://www.R-project.org/> (2015).



Extended Data Figure 1 | Diagram depicting the experimental design.

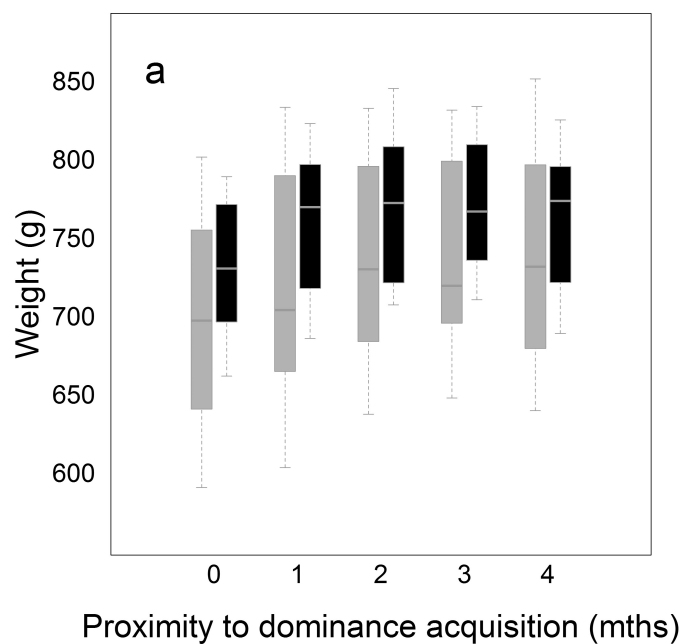
Juvenile experiments were conducted from 15 December 2010 to 19 August 2012, and adult experiments from 28 March 2011 to 20 July 2013. Each horizontal line represents longitudinal weight data collected from an experimental group. Thick orange lines represent unfed, challenged individuals and blue lines represent fed challengers. Thick green lines represent control individuals, which were animals of the same sex and age-range from the same population over the same period (2010–2013).

Red boxes indicate the 3-month experimental windows of food supplementation, which spanned different periods for different dyads (allowing us to disentangle experimental effects from environmental and seasonal effects on weight) and, for the adult experiment, occurred any time between 310 and 870 days of age. F, female; M, male. Note that the *x* axis is not drawn to scale, to facilitate comparison of the design between the juvenile and adult cohorts. The meerkat icon was downloaded from PhyloPic: <http://phylopic.org>, with credit to M. Keesey.

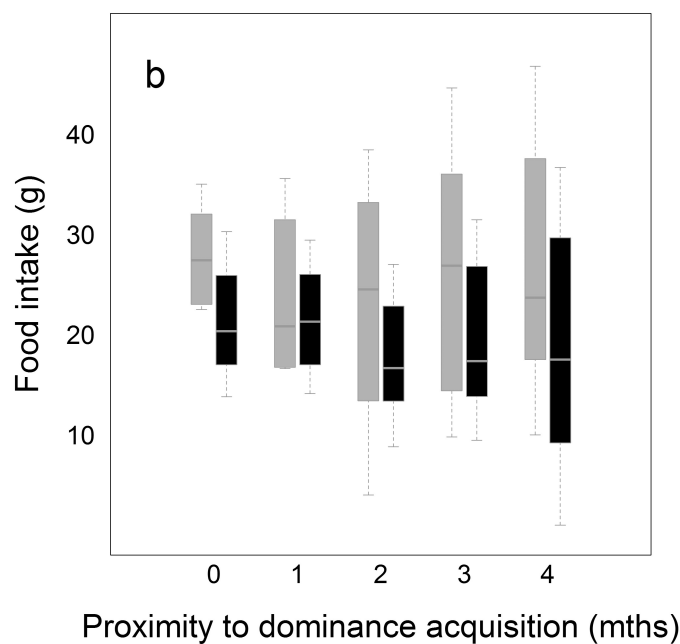


Extended Data Figure 2 | Relationship between the growth of the challenged individual and the growth of its fed challenger. a, Juveniles; b, adults. Thirty-two juvenile and 17 adult experimental pairs were included. Growth was calculated as the individual weight difference

between the start and mid-point of the experiment. Dots show the raw values (grey for females, black for males). The dotted line shows the predicted values of the linear model (results presented in Extended Data Table 1) and s.d. of the predicted values are delineated by shaded areas.



Extended Data Figure 3 | Changes in weight and food intake (average morning weight gain) in new dominant females (grey boxes, $n = 42$) and males (black boxes, $n = 30$). a, Weight; b, food intake. Boxplots show the raw values, averaged for each individual during the month preceding dominance



acquisition (labelled '0'), as well as during the first, second, third and fourth months' post-dominance acquisition (respectively labelled '1', '2', '3' and '4'). Whiskers show all data points that are no further away from the box than half the interquartile range.

Extended Data Table 1 | Results of linear models investigating the relationship between the growth of challenged individuals and their fed challengers in juveniles and adults

Variable	Est.	SE	DF	F-value	P-value
JUVENILES					
Growth of fed challenger (g)	1.068	0.17	27	39.43	<10 ⁻⁴
Sex	-14.178	8.50	27	2.78	0.107
Age	0.726	0.94	27	0.59	0.448
Rainfall	0.012	0.09	27	0.02	0.897
ADULTS					
Growth of fed challenger (g)	0.916	0.24	12	14.72	0.002
Sex	6.143	13.99	12	0.19	0.668
Age	-0.164	0.06	12	7.16	0.020
Rainfall	0.205	0.08	12	7.19	0.020

The response variable is the growth of the challenged individual, calculated as the individual weight difference (g) between the start and mid-point of the experiment. The juvenile model includes 12 females and 20 males; the value of the model adjusted R^2 is 0.65. The adult model includes 8 females and 9 males; the value of the model adjusted R^2 is 0.61. Est., estimate.

Extended Data Table 2 | Results of the post hoc paired t-tests investigating temporal changes in weight following dominance acquisition

		Proximity from dominance acquisition (months)			
		df=47 for all tests			
		1	2	3	4
Proximity from dominance acquisition (months)	0	t=4.34, p<0.001	t=5.83, p<10 ⁻⁴	t=7.28, p<10 ⁻⁴	t=5.09, p<10 ⁻⁴
	1	—	t=3.52, p<0.001	t=3.94, p=0.003	t=2.63, p=0.115
	2	—	—	t=0.90, p=1.000	t=0.14, p=1.000
	3	—	—	—	t=0.78, p=1.000
	4	—	—	—	—

Pairwise comparison tests were conducted after the repeated-measures analysis of variance to compare within-individual changes in weight between the month preceding dominance acquisition (labelled '0') and the 4 months (labelled '1'–'4') following dominance acquisition, as well as between each of the 4 months' post-dominance acquisition. A Bonferroni correction was applied to correct for multiple testing.

Extended Data Table 3 | Results of the linear model investigating changes in body weight within 150 days following dominance acquisition in relation to absolute weight difference with the heaviest same-sex subordinate

Variable	Est.	SE	DF	F-value	p-value
Age at dominance acquisition (days)	-0.030	0.02	36	2.59	0.117
Sex (reference: female)	-5.541	28.75	36	0.04	0.848
Rainfall (mm)	-0.270	0.11	36	5.65	0.023
Seasonality	5.425	11.04	36	0.24	0.626
Weight gap with main rival (g)	-0.758	0.27	36	7.69	0.009
Sex : weight gap with main rival	0.597	0.39	36	2.29	0.139

This analysis includes 25 females and 20 males. The value of the model adjusted R^2 is 0.21.

How sexual selection can drive the evolution of costly sperm ornamentation

Stefan Lüpold^{1,2*}, Mollie K. Manier^{1,3}, Nalini Puniamoorthy^{1,4}, Christopher Schoffl¹, William T. Starmer¹, Shannon H. Buckley Luepold¹, John M. Belote¹ & Scott Pitnick^{1*}

Post-copulatory sexual selection (PSS), fuelled by female promiscuity, is credited with the rapid evolution of sperm quality traits across diverse taxa¹. Yet, our understanding of the adaptive significance of sperm ornaments and the cryptic female preferences driving their evolution is extremely limited^{1,2}. Here we review the evolutionary allometry of exaggerated sexual traits (for example, antlers, horns, tail feathers, mandibles and dewlaps), show that the giant sperm of some *Drosophila* species are possibly the most extreme ornaments^{3,4} in all of nature and demonstrate how their existence challenges theories explaining the intensity of sexual selection, mating-system evolution and the fundamental nature of sex differences^{5–9}. We also combine quantitative genetic analyses of interacting sex-specific traits in *D. melanogaster* with comparative analyses of the condition dependence of male and female reproductive potential across species with varying ornament size to reveal complex dynamics that may underlie sperm-length evolution. Our results suggest that producing few gigantic sperm evolved by (1) Fisherian runaway selection mediated by genetic correlations between sperm length, the female preference for long sperm and female mating frequency, and (2) longer sperm increasing the indirect benefits to females. Our results also suggest that the developmental integration of sperm quality and quantity renders post-copulatory sexual selection on ejaculates unlikely to treat male–male competition and female choice as discrete processes.

Across animals, the sex competing more intensely for mates has evolved more elaborate ornaments and/or weapons functioning in mate acquisition¹⁰. Because these secondary sexual traits are typically costly, their growth is highly responsive to physiological correlates of their bearer's nutritional state¹¹, which is influenced by both genes and environment. Such condition-dependent expression¹² is a foundation of sexual selection theory and indicator models (for example, 'good genes' and 'handicap') of mate choice^{10,13}. It also explains why ornament size generally increases disproportionately with body size ('positive allometry', slope of log–log regression > 1.0) within and among species¹⁴, with typical among-species slopes of 1.4–3.8 (Extended Data Table 1).

The relative intensity of competition for mates is often heavily influenced by the ratio of reproductively available males and females⁶, which itself is influenced by their relative reproductive potential⁹. Males frequently have a greater reproductive potential due to lower production costs of sperm relative to eggs⁵ and typically smaller paternal investment in offspring^{7,9}. These sexual disparities and their link to sexual selection provides another foundation of sexual selection theory and explains why males commonly are the more aggressive and/or more ornamented sex^{5,7–10}. Broad theoretical and empirical work indicates that stronger premating sexual selection correlates with more extreme ornamentation and greater sex differences in reproductive potential^{9,10}.

Since both sexes are promiscuous in most species, intrasexual competition and intersexual choice can continue after mating through

sperm competition¹⁵ and cryptic female choice². The best-known adaptation to post-copulatory sexual selection (PSS) is the production of copious sperm. More sperm should nearly always enhance competitive fertilization success, thus explaining the widespread positive correlation between relative testis size and sperm competition risk¹⁵. Taxa with this adaptation will tend to exhibit positive covariation between the strength of PSS and sexual disparity in reproductive potential, similar to the pattern for premating sexual selection.

A theoretical conundrum arises, however, when considering that PSS also selects for longer sperm in *Drosophila*^{3,16–18} and numerous other taxa¹. Because sperm length competes locally for resources with sperm number owing to their spatial and temporal co-occurrence within the developmental environment of the testes, the two traits are relatively constrained to evolutionarily trade off against one another¹⁹. Across *Drosophila* species, sperm length displays strong negative correlation with both the number of sperm manufactured (slope = –0.97, $R^2 = 0.55$) and ejaculated (slope = –1.56, $R^2 = 0.90$)²⁰. Consequently, species with gigantic sperm (and particularly intense PSS) exhibit the least sex difference in reproductive potential⁴. For example, *D. bifurca* has 5.8-cm-long sperm, and only a few times more sperm than eggs are produced in the population⁴. Because sexual selection theory predicts the weakest sexual selection for such species (see above), this phenomenon was coined the 'big-sperm paradox'⁴.

To better characterize this paradox, we first examined the evolutionary allometry of sperm length and egg volume across all *Drosophila* species that had reports for both traits in the literature ($n = 46$ species; Extended Data Table 2 and Extended Data Fig. 1) using phylogenetic reduced major-axis (RMA) regressions. The slope of the sperm-length allometry was 5.52 (Fig. 1a; $P < 0.0001$, $\lambda = 1.0$), which is approximately twofold greater than slopes for nearly all other sexually selected traits previously studied (Fig. 2; Supplementary Tables 1–3; Extended Data Figs 2 and 3). In sharp contrast, linearized egg size was negatively allometric, albeit not significantly so (Fig. 1b; slope = 0.84, $P = 0.19$, $\lambda = 1.00$). We further examined all available data on ovariole number for this set of species as an index of the number of eggs produced²¹ and found it to exhibit positive allometry ($n = 35$, slope = 2.63, $P < 0.0001$, $\lambda = 0.99$). Finally, egg volume declined as ovariole number increased in a phylogenetic regression controlled for body size ($n = 35$, $r = -0.69$, $P < 0.0001$; thorax length: $r = 0.77$, $P < 0.0001$; $\lambda < 0.0001^{1.00,0.02}$). That larger-bodied species produce fewer, longer sperm, yet more eggs, reinforces the big-sperm paradox by further limiting the number of sperm competing for each egg⁴ and hence the predicted intensity of PSS on sperm quality⁹. Bjork and Pitnick⁴ showed that, contrary to theoretical prediction, the 'opportunity for sexual selection', which is the standardized intra-sexual variance in the number of offspring produced and expresses the maximum potential strength of sexual selection²², did not decline with increasing sperm length. Moreover, the female-specific opportunity for sexual selection increased with sperm length

¹Center for Reproductive Evolution, Department of Biology, Syracuse University, 107 College Place, Syracuse, New York 13244-1270, USA. ²Department of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland. ³Department of Biological Sciences, The George Washington University, 800 22nd St. NW, Suite 6000, Washington DC 20052, USA. ⁴Department of Biological Sciences, National University of Singapore, 14 Science Drive, SG 117543, Singapore.

*These authors contributed equally to this work.

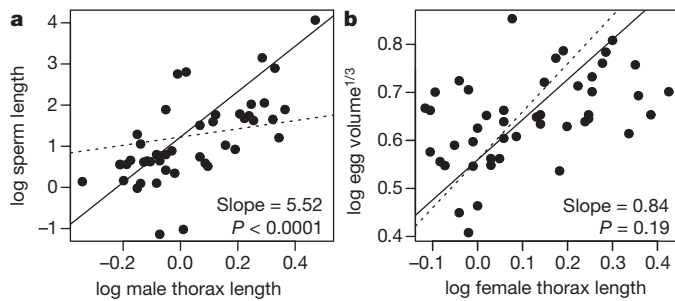


Figure 1 | Allometry of sperm length and egg volume. **a, b,** Interspecific allometric relationships of sperm length (**a**; slope = 5.52, $P < 0.0001$, $\lambda = 1.00$) and egg volume (**b**; slope = 0.84, $P = 0.19$, $\lambda = 1.00$) for 46 *Drosophila* species. Egg volume was linearized by taking the cube root for geometric scaling with thorax length²¹ and consistent dimensionality with sperm length. Egg length yielded identical results. Dotted lines represent isometry (slope = 1.0).

($R^2 = 0.994$)⁴. However, Bjork and Pitnick⁴ were unable to explain these patterns despite the ratio of sperm to eggs approaching parity.

Achieving a resolution to the big-sperm paradox requires explaining the mechanism(s) by which a stronger female preference compensates for the theoretically predicted (but not realized⁴) intrinsic decline in the strength of PSS resulting from reduced sperm numbers with increasing investment per sperm. A resolution should also discern how females benefit from their preference for longer sperm. The length of the female's primary sperm-storage organ, the seminal receptacle (SR), co-diversifies with sperm length in *Drosophila*²³ and numerous other taxa¹ and has been demonstrated to be the proximate basis of a cryptic female preference for sperm length. Specifically, longer sperm are superior at displacing, and resisting displacement by, shorter competitor sperm within the SR^{3,16–18}, and longer SRs drive sperm-length evolution by enhancing this competitive advantage³. Because there are substantive developmental and longevity costs associated with longer SRs¹⁸, SR length is more likely to evolutionarily increase if these costs are compensated for by direct and/or indirect benefits accrued by biasing fertilization in favour of longer sperm. Although *Drosophila* sperm have been shown to contribute no direct benefits to the female or her offspring^{24,25}, indirect benefits postulated to explain the evolution of premating female preferences may similarly explain cryptic postmating female preferences².

We first investigated whether Fisherian runaway sexual selection could provide a countervailing mechanism for the intrinsic decline in the strength of selection predicted to accompany increases in sperm length. We conducted an intraspecific test of an essential prediction of this hypothesis—a positive genetic correlation between SR and sperm length—using a well-replicated diallel breeding design between ten *D. melanogaster* isogenic lines and evaluating the genetic architecture underlying trait variation (see Methods and also ref. 26). We found a highly significant, positive genetic correlation between sperm and SR length (Table 1), which would theoretically serve to drive sperm-length evolution as SR length evolves (and vice versa). Importantly, increases in SR length would further intensify directional selection on sperm length, as SR length was negatively genetically correlated with female remating interval and positively correlated with the time interval between insemination and active female ejection of excess last-male and displaced resident sperm from the reproductive tract (Table 1). Faster remating enhances PSS, and later sperm ejection prolongs direct competition between sperm for limited storage space and affords longer sperm greater opportunity to exert their superior competitiveness²⁶ (also note the positive genetic correlation between SR length and the proportion of resident sperm displaced; Table 1).

We next explored the potential for females to accrue indirect (genetic) benefits by virtue of sperm length serving as a reliable indicator of male quality. We compared *D. melanogaster* reared in benign

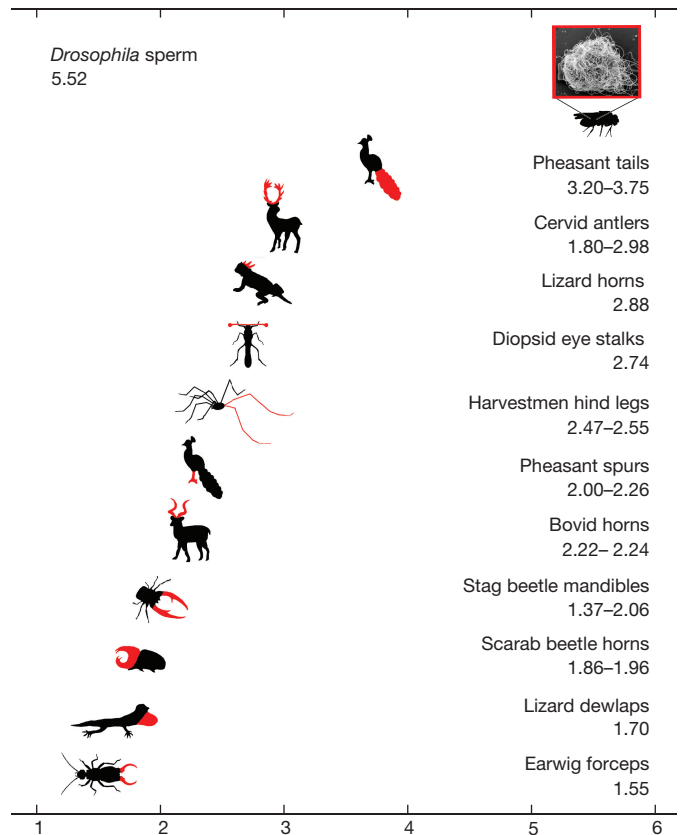


Figure 2 | Evolutionary allometry of *Drosophila* sperm length in comparison with other, classic examples of sexually selected traits.

Values are interspecific allometric slopes. Detailed statistics and data sources are listed in Extended Data Table 1. Inset, scanning electron micrograph of single, 58.3-mm-long sperm of *D. bifurca*. Image courtesy of Romano Dallai.

and stressful developmental environments within a quantitative genetic framework to assess the sensitivity of sperm length to the nutritional history and the physiological condition of males^{11–13}. Sperm length was highly heritable (Table 1) but not condition-dependent (linear mixed-effects model controlling for genetic background of 45 nuclear genotypes: $t = -0.57$, $P = 0.58$; Extended Data Fig. 4). At face value, this result refutes all indicator models as an explanation for SR-length evolution. Nevertheless, because of the strong negative evolutionary relationship between sperm length and number in *Drosophila*²⁰, sperm-length evolution may be mediated by its influence on the condition dependence of sperm number. We thus investigated seven *Drosophila* species varying in body sizes, sperm lengths and egg volumes (Extended Data Table 2; Extended Data Fig. 5). Rearing each under varying larval densities, we produced a range of adult body sizes as a proxy for condition^{12,13,27}, as previous studies employing a similar approach with *Drosophila* have demonstrated positive associations between male body size and fitness²⁸. These adults were assayed for reproductive potential with no reproductive competition and *ad libitum* access to mates, food and oviposition substrate. We then examined the strength and slope of the within-species, sex-specific relationships between body condition and reproductive potential (see Methods) to test the prediction that male reproductive potential becomes increasingly condition-dependent as sperm length increases.

Male reproductive potential increased with condition in all species (Extended Data Fig. 6a–g), although not significantly so in *D. arizonae* with the shortest sperm (Extended Data Fig. 6a; $r = 0.36$, $P = 0.11$; all other species: $r \geq 0.49$, $P \leq 0.01$; Extended Data Table 3 and Extended Data Fig. 5a, c). *Drosophila bifurca*, with the longest sperm, exhibited the strongest relationship ($r = 0.93$, $P < 0.0001$;

Table 1 | Bootstrapped genetic correlations, phenotypic correlations and heritabilities in sperm length, female morphology and traits related to sperm storage and use, based on means within diallel crosses ($n = 90$)

	Sperm length	SR length	Remating day	Eject time	Prop. sperm displaced
Sperm length	0.265 ± 0.107*	0.683 ± 0.297*	-0.589 ± 0.594	-0.431 ± 0.414	0.923 ± 1.510
SR length	0.369 ± 0.081*	0.192 ± 0.048*	-0.793 ± 0.285*	0.423 ± 0.210*	1.051 ± 0.394*
Remating day	-0.079 ± 0.105	-0.337 ± 0.071*	0.103 ± 0.047*	-0.819 ± 0.375*	-0.301 ± 0.377
Eject time	0.045 ± 0.098	0.116 ± 0.071	-0.125 ± 0.075	0.142 ± 0.074*	0.847 ± 0.292*
Prop. sperm displaced	0.160 ± 0.111	0.129 ± 0.070†	-0.024 ± 0.076	0.337 ± 0.067*	0.090 ± 0.040*

Additive genetic correlations ($r_A \pm$ s.e.) are given above the diagonal, heritabilities ($h^2 \pm$ s.e.; boldface) on the diagonal and phenotypic (Pearson's) correlations ($r \pm$ s.e.) below the diagonal. Prop., proportion.

*Significant correlations at $\alpha < 0.05$.

† $P = 0.065$.

Extended Data Figs 5b, d and 6g; Extended Data Table 3). Female reproductive potential similarly increased with body size in all species, albeit non-significantly in *D. arizonae* and *D. hydei* ($r \leq 0.08$, $P \geq 0.65$; all other species: $r \geq 0.45$, $P \leq 0.01$; Extended Data Fig. 6h–n; Extended Data Table 3). Note that *D. arizonae* (Extended Data Fig. 6h) has the smallest eggs and *D. hydei* (Extended Data Fig. 6m) has medium-sized eggs; *D. melanogaster* showed the strongest relationship (Extended Data Fig. 6i), also with medium-sized eggs (Extended Data Table 2).

Next, we combined these intraspecific relationships for all seven species into comparative analyses to determine how much of the among-species variation in the condition dependence of sex-specific reproductive potential is explained by variation in gamete size (Fig. 3). In phylogenetic regressions, the male reproductive potential became increasingly condition-dependent as sperm length increased ($r = 0.82$, $P = 0.02$, $\lambda < 0.0001^{1.0,0.04}$; Fig. 3a), with the standardized slopes also becoming steeper ($r = 0.94$, $P = 0.002$, $\lambda = 1.0^{0.09,1.00}$; Fig. 3b). Hence, males of any condition can produce and inseminate many 'cheap' sperm, but only high-quality males have the available resources to produce abundant 'expensive' sperm. In striking contrast, producing larger eggs did not increase the condition dependence of the reproductive potential in females ($r = 0.51$, $P = 0.24$, $\lambda < 0.0001^{1.0,0.17}$; Fig. 3c), nor

did the intraspecific slopes become steeper as egg volume increased ($r = 0.66$, $P = 0.11$, $\lambda < 0.0001^{1.0,0.11}$; Fig. 3d). Hence, investment per gamete underlies interspecific variation in the condition dependence of reproductive potential for males but not females.

Our findings offer a possible resolution to the big-sperm paradox by revealing an interacting combination of trait covariance and mating-system characteristics antithetical to the weakening of the sexual selection intensity as sperm length increases. Given the substantial costs of producing long sperm^{20,29}, it is unclear how this trait has evaded the theoretically predicted development of condition dependence found for other costly sexual characters¹³. Nevertheless, the intimate developmental association between sperm length and number renders the latter trait a surrogate indicator of correlated condition. Smaller (poor-quality) males pay higher costs for the same increase in trait size^{11,30}, making the production of plentiful long sperm an intrinsically 'unfakeable' trait. Females of species with longer SRs remate more frequently, owing to both a negative genetic correlation between the two traits and faster sperm depletion when receiving smaller ejaculates. In *D. bifurca* and other species with very long sperm, females typically mate with several males each day⁴, which may explain the previously observed, strong positive relationship between sperm length and the female-specific opportunity for sexual selection⁴. What is perhaps most critical to our understanding of sperm-length evolution is that only males in good condition can produce sufficient sperm to capitalize on the increased mating opportunities, with females consequently receiving indirect genetic benefits. These results reveal a novel component to our understanding of the operation of sexual selection: the intensity of selection on female preferences can remain strong owing to within-population variance in male reproductive potential, even when sex-specific mean reproductive potentials and the operational sex ratio approach unity.

By experimentally manipulating sperm length and number in *D. melanogaster*, both traits were previously found to contribute to competitive fertilization success, with the relative fitness contribution of sperm length increasing as sperm numbers decreased¹⁶. Here we further demonstrate the non-independence of selection on sperm quantity and quality, and hence the false dichotomy of sperm competition and cryptic female choice as forces shaping the evolution of sperm form. For many species, what may matter most in PSS is not simply transferring the most sperm or the best sperm, but rather the greatest number of sperm that are designed to survive and compete best given the specific female reproductive environment.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 January 2016; accepted 13 April 2016.

1. Pitnick, S., Hosken, D. J. & Birkhead, T. R. in *Sperm Biology: An Evolutionary Perspective* (eds Birkhead, T. R., Hosken, D. J. & Pitnick, S.) 69–149 (Academic Press, 2009).
2. Eberhard, W. G. *Female Control: Sexual Selection by Cryptic Female Choice*. (Princeton University Press, 1996).
3. Miller, G. T. & Pitnick, S. Sperm-female coevolution in *Drosophila*. *Science* **298**, 1230–1233 (2002).

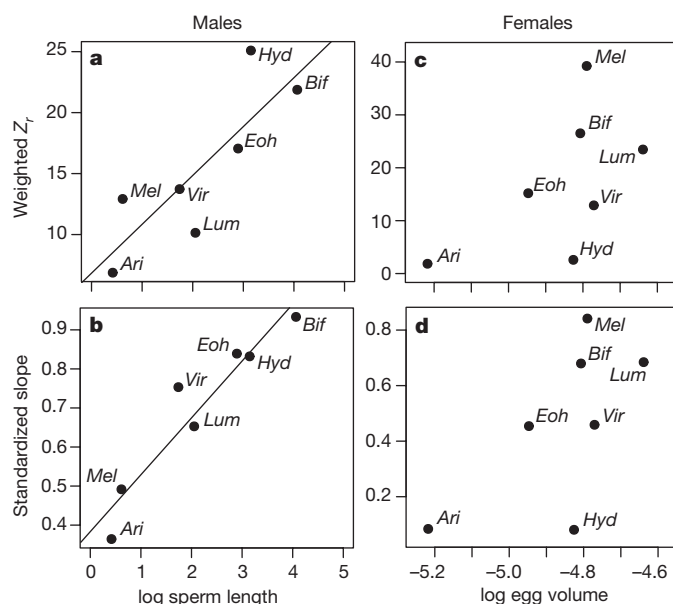


Figure 3 | Comparison of intraspecific condition dependence of sperm length and egg volume across seven *Drosophila* species. a–d, For males, sperm length predicts across seven *Drosophila* species the degree to which reproductive potential correlates with body size (a) and the slope of the relationship (b), whereas egg size does not significantly predict either the strength (c) or the slope (d) of this relationship in females (see Extended Data Fig. 6). The weighted Z_r values reflect the correlation coefficients of intraspecific relationships between reproductive potential and body size for either males (a) or females (c; for details see Methods). Figures are not controlled for phylogeny. Ari, *D. arizonae*; Mel, *D. melanogaster*; Vir, *D. virilis*; Lum, *D. lummei*; Eoh, *D. eohydei*; Hyd, *D. hydei*; Bif, *D. bifurca*.

4. Bjork, A. & Pitnick, S. Intensity of sexual selection along the anisogamy-isogamy continuum. *Nature* **441**, 742–745 (2006).
5. Bateman, A. J. Intra-sexual selection in *Drosophila*. *Heredity* **2**, 349–368 (1948).
6. Parker, G. A., Baker, R. R. & Smith, V. G. The origin and evolution of gamete dimorphism and the male-female phenomenon. *J. Theor. Biol.* **36**, 529–553 (1972).
7. Trivers, R. L. in *Sexual Selection and the Descent of Man 1871–1971* (ed. Campbell, B.) 136–179 (Aldine-Atherton, 1972).
8. Emlen, S. T. & Oring, L. W. Ecology, sexual selection, and the evolution of mating systems. *Science* **197**, 215–223 (1977).
9. Clutton-Brock, T. H. & Parker, G. A. Potential reproductive rates and the operation of sexual selection. *Q. Rev. Biol.* **67**, 437–456 (1992).
10. Andersson, M. *Sexual Selection*. (Princeton University Press, 1994).
11. Emlen, D. J., Warren, I. A., Johns, A., Dworkin, I. & Lavine, L. C. A mechanism of extreme growth and reliable signaling in sexually selected ornaments and weapons. *Science* **337**, 860–864 (2012).
12. Bonduriansky, R. *et al.* Differential effects of genetic vs. environmental quality in *Drosophila melanogaster* suggest multiple forms of condition dependence. *Ecol. Lett.* **18**, 317–326 (2015).
13. Rowe, L. & Houle, D. The lek paradox and the capture of genetic variance by condition dependent traits. *Proc. R. Soc. Lond. B* **263**, 1415–1421 (1996).
14. Kodric-Brown, A., Sibly, R. M. & Brown, J. H. The allometry of ornaments and weapons. *Proc. Natl Acad. Sci. USA* **103**, 8733–8738 (2006).
15. Parker, G. A. & Pizzari, T. Sperm competition and ejaculate economics. *Biol. Rev. Camb. Philos. Soc.* **85**, 897–934 (2010).
16. Pattarini, J. M., Starmer, W. T., Bjork, A. & Pitnick, S. Mechanisms underlying the sperm quality advantage in *Drosophila melanogaster*. *Evolution* **60**, 2064–2080 (2006).
17. Lüpold, S. *et al.* How multivariate ejaculate traits determine competitive fertilization success in *Drosophila melanogaster*. *Curr. Biol.* **22**, 1667–1672 (2012).
18. Miller, G. T. & Pitnick, S. Functional significance of seminal receptacle length in *Drosophila melanogaster*. *J. Evol. Biol.* **16**, 114–126 (2003).
19. Nijhout, H. F. & Emlen, D. J. Competition among body parts in the development and evolution of insect morphology. *Proc. Natl Acad. Sci. USA* **95**, 3685–3689 (1998).
20. Pitnick, S. Investment in testes and the cost of making long sperm in *Drosophila*. *Am. Nat.* **148**, 57–80 (1996).
21. Starmer, W. T. *et al.* in *Evolutionary Biology* (eds Macintyre, R. J. & Clegg, M. T.) 139–171 (Springer, 2003).
22. Wade, M. J. Sexual selection and variance in reproductive success. *Am. Nat.* **114**, 742–747 (1979).
23. Pitnick, S. S., Markow, T. A. & Spicer, G. S. Evolution of multiple kinds of female sperm-storage organs in *Drosophila*. *Evolution* **53**, 1804–1822 (1999).
24. Karr, T. L. & Pitnick, S. The ins and outs of fertilization. *Nature* **379**, 405–406 (1996).
25. Pitnick, S., Spicer, G. S. & Markow, T. A. Phylogenetic examination of female incorporation of ejaculates in *Drosophila*. *Evolution* **51**, 833–845 (1997).
26. Lüpold, S. *et al.* Female mediation of competitive fertilization success in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **110**, 10693–10698 (2013).
27. Bonduriansky, R. & Day, T. Nongenetic inheritance and the evolution of costly female preference. *J. Evol. Biol.* **26**, 76–87 (2013).
28. Partridge, L. & Farquhar, M. Lifetime mating success of male fruitflies (*Drosophila melanogaster*) is related to their size. *Anim. Behav.* **31**, 871–877 (1983).
29. Pitnick, S., Markow, T. A. & Spicer, G. S. Delayed male maturity is a cost of producing large sperm in *Drosophila*. *Proc. Natl Acad. Sci. USA* **92**, 10614–10618 (1995).
30. Pitnick, S. & Markow, T. A. Large-male advantages associated with costs of sperm production in *Drosophila hydei*, a species with giant sperm. *Proc. Natl Acad. Sci. USA* **91**, 9277–9281 (1994).

Supplementary Information is available in the online version of the paper.

Acknowledgements The authors thank B. Reil for technical assistance and S. Dorus for helpful comments on the manuscript. Financial support for this research was provided by the National Science Foundation (grants DEB-9806649 to S.P. and DEB-1145965 to S.P., S.L., M.K.M. and J.M.B.), the Swiss National Science Foundation (Fellowships PA00P3_134191 and PZ00P3_154767 to S.L.), the National University of Singapore (Overseas Postdoctoral Fellowship to N.P.) and a generous gift from Mike and Jane Weeden to Syracuse University.

Author Contributions S.P. and S.L. conceived the research. S.P. and C.S. performed the reproductive potential experiments. S.P., C.S. and W.T.S. collected data for sperm and egg production allometry. S.L., S.P., M.K.M. and J.M.B. performed the male–female trait genetic covariance experiments. S.P., S.L., N.P. and S.H.B.L. performed the sperm length condition dependence experiment. S.L. and W.T.S. performed all statistical analyses. S.P. and S.L. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.P. (sspitnic@syr.edu).

METHODS

No statistical methods were used to predetermine sample size. Flies were randomly assigned to experimental treatments. All measurements and counts were conducted blind to treatment and to values of other traits and outcomes in mating experiments.

Experimental material. Condition dependence of sex-specific reproductive potential was assayed using strains of *D. eohydei* (15085-1631.0), *D. bifurca* (15085-1621.0), *D. virilis* (15010-1051.0), and *D. lummei* (15010-1011.1) obtained from the National Drosophila Species Center, San Diego, California. *Drosophila hydei* was collected at the South Coast Agricultural Research Station, California by J. Graves; *D. melanogaster* was collected in Napa Valley, California by D. Begun; *D. arizonae* was collected in the Superstition Mountains, Arizona by T. A. Markow. All species were cultured on cornmeal–agar–molasses medium under uncrowded conditions and 1:1 sex ratio in 200-ml bottles with live yeast at $24 \pm 1^\circ\text{C}$ and a 12-hour light/dark photoperiodic cycle.

Quantitative genetic analyses of sperm and female reproductive tract morphology, sperm handling and sperm competition outcomes was performed with genetically transformed LH_m populations of *D. melanogaster* that express a protamine labelled with either green (GFP) or red fluorescent protein (RFP) in sperm heads³¹. All experimental flies derived from isogenic lines³² ('isolines') of the respective GFP and RFP populations, following 15 generations of full-sibling inbreeding (theoretical inbreeding coefficient = 0.96)³³.

Evolutionary allometry of sperm and egg size. Sperm length, egg volume, ovariule number and the sex-specific thorax length data for 46 species were obtained from the literature^{21,29,34}, with novel data (except ovariule number) obtained for ten additional species using identical methods (Extended Data Table 2). *Drosophila ficusphila* was excluded from the analyses including ovariule number due to being an extreme outlier (13 compared to 22.6–52.86 ovariules in all other species; Extended Data Table 2).

Evolutionary allometry of exaggerated, sexually selected traits from different taxa. For comparison of the allometric slope of *Drosophila* sperm with slopes of other sexual traits that are widely considered to be exaggerated due to intra- or intersexual selection, we obtained interspecific allometric slopes or comparative data sets permitting such analyses from the literature for a range of classic examples^{14,35–38} (Extended Data Table 1; Supplementary Tables 1–3). Reported allometric slopes were not usually controlled for phylogeny and could not always be reanalysed because data sets were not provided, but where possible, we reanalysed them by incorporation of a molecular phylogeny (Extended Data Figs 2 and 3; Supplementary Table 3). Since all phylogenies were reconstructed from published figures without branch length information or were combined from different molecular trees, we used equal branch lengths in all taxa. Based on slope comparisons with and without phylogenetic control, however, the lack of such control did not have a major impact on the interspecific slopes. Within these constraints, precise slope estimates should be used with care.

Condition dependence of sperm length. Using the same isolines as the quantitative genetic analyses (see below) but in a half-diallel instead of diallel cross design (that is, $n = 45$), 40 newly-hatched larvae of each cross were transferred to a rearing vial with regular fly medium (see above) and another 40 larvae to a vial with 75% less yeast in the medium and only half the amount of medium in the vial. Larvae were randomly assigned to rearing treatments. Following development under these benign and moderately stressful conditions, respectively, five random males of each cross and rearing treatment were aged for at least a week before measuring their thorax length and the length of five sperm per male.

Condition dependence of reproductive potential. For all seven species, variation in body size was generated by transferring first-instar larvae randomly to culture vials at three different densities: 25, 75, and 150 larvae per 8-dram vial containing 8 ml of medium. Virgin flies were then collected on the day of eclosion and thorax length, a reliable index of total dry mass³⁹, was recorded. Focal males and females were selected to represent the entire size distribution, with each fly then isolated within a vial containing medium and live yeast and transferred to a fresh vial every three days until reaching two days post-reproductive maturity, the age of which varies between sexes and among species²⁹. All virgin males and females used as mates of focal flies were derived from population bottles.

The reproductive potential of each focal male ($n = 15$ –27 per species) was assayed by placing it with eight randomly assigned virgin females in a plastic 200 ml bottle that was inverted over a small Petri dish containing medium and live yeast. Every 24 h, across four successive days, the male was removed and transferred to a new bottle containing eight virgin females. Because males could exhibit size-related variation in the number of mature sperm stored in the seminal vesicles at the start of the experiment, the eight females from day 1 were discarded. The 24 females from days 2–4 were provided with fresh oviposition plates daily until the production of offspring ceased (that is, no eggs hatched). Oviposition plates

were stored at 25°C and the number of larvae hatching on each plate was counted after 48 h. All larvae produced by the 24 females exposed to each male were summed as a measure of that male's reproductive potential.

Female reproductive potential was assayed in a manner similar to males, except that each focal female ($n = 25$ –36 per species) was placed with three randomly assigned virgin males in a vial containing medium and live yeast. Each focal female was transferred to a fresh vial with three new virgin males every 24 h across four successive days. The day 1 vial was discarded to control for variation among females in the number of mature oocytes at the start of the experiment. All eggs laid by each female from days 2–4 were summed as a measure of that female's reproductive potential.

Quantitative genetic analyses of female preference, male ornament and associated characters. To vary the female genetic background, single pairs of virgin males and females of ten different RFP isolines were crossed in all non-self combinations (that is, 90 diallel crosses with 45 different nuclear genotypes, all independent of the RFP standard competitor male²⁶). In each of two blocks separated by two generations, we assayed three random F_1 females from each of three separate male–female pairs per cross (that is, 90 crosses \times 2 blocks \times 3 families \times 3 females = 1,620 females). All virgin flies were aged for three days before their first mating. All experimental males were F_1 progeny from crosses among a single pair of isolines with either GFP- or RFP-tagged sperm.

Using a double-mating design, reproductive outcomes were quantified immediately after female sperm ejection (that is, <5 h after mating and before the first egg has entered the bursa for fertilization) following the second mating, which we have shown repeatedly to directly predict paternity shares among competing males over the three subsequent days of oviposition^{17,26,31}. Each female was mated with a virgin GFP male and, two days later, with a virgin RFP male, with additional 6-h remating opportunities on days 3–4 for any refractory females. Each male was used for only one mating. Following all matings with a second male, we used established protocols to quantify (i) copulation duration, (ii) the number of resident first-male sperm at the time of remating, (iii) time until female ejection of excess second-male and displaced first-male sperm, (iv) the number of displaced first-male sperm, the number of second-male sperm (v) transferred and (vi) ejected, (vii) the proportion of each male's sperm ejected, (viii) the distribution of both competitors' sperm, respectively, across the different organs of the female reproductive tract (that is, bursa copulatrix, SR, and paired spermathecae) and (ix) the proportional representation of sperm derived from the first (S_1) or second male (S_2) in each respective location (for example, the SR, which is the primary source of sperm for fertilization³¹) and in the entirety of the female reproductive tract. For one random female of each family (that is, six females per cross), we additionally measured the length of the thorax and the $\text{SR}^{17,26,31}$.

Statistical analyses. All analyses were performed using the statistical package R version 3.0.2 (R Development Core Team 2013) and SAS v9.3 (SAS Institute 2011).

Evolutionary allometry of sperm and eggs. We used phylogenetically controlled reduced major-axis regressions (phyl.RMA in R package phytools). For these analyses, additional species (that is, *D. mettleri*, *D. pachea*, *D. subpalustris*, *D. rhopala* and *D. suzukii*) were added to the van der Linde *et al.*⁴⁰ phylogeny based on other molecular phylogenies^{29,41} (Extended Data Fig. 1). We linearized egg volume by the cube root for consistent dimensionality with female thorax length and sperm length²². For comparison, however, we also used egg length, the allometric slope of which was identical to linearized egg volume up to the third decimal point ($b = 0.836$ compared to 0.835).

Evolutionary allometry of exaggerated, sexually selected traits from different taxa. Wherever data and corresponding phylogenies were available, we analysed them using phyl.RMA as for *Drosophila* gametes. For direct comparison between taxa and/or traits, we adjusted all data to equal dimensionality (that is, cube-rooting mass variables or square-rooting area variables) to ensure that isometry was at a slope of 1. All analyses were confirmed to exhibit a significant association between the two traits compared in phylogenetic least-squares regressions before calculating phylogenetic RMA slopes.

Condition dependence of sperm length. Treatment effects on sperm length were analysed in linear mixed-effects models controlling for the genetic background of sires and dams and their interaction as random effects. For comparison, we repeated these analyses on the thorax length of the same males.

Condition dependence of reproductive potential. For each of the seven species, regression analyses were used to examine the relationship between either the total number of progeny produced and male size (that is, thorax length) or the total number of eggs laid and female size. For these relationships, we calculated the intraspecific correlation coefficients, r , which represent their strength and direction, as well as the standardized slopes, for use in subsequent comparative analyses. A Bartlett's test of homogeneity of variances confirmed no differences among the seven species in the coefficient of thorax length for males ($K^2 = 9.92$, $P = 0.13$).

Although there was a marginally significant difference for females ($K^2 = 12.67$, $P = 0.05$), this was primarily attributable to a greater standard deviation in female thorax length in *D. hydei* (Extended Data Fig. 7; a Bartlett's test revealed no significant difference among the remaining species when *D. hydei* was excluded: $n = 6$, $K^2 = 4.38$, $P = 0.50$).

To compare the degree of intraspecific condition dependence among species, we converted the correlation coefficients, r , of the intraspecific regressions using Fisher's transformation and weighted them by sample size to obtain a weighted Z_r for each species⁴². Comparative relationships between weighted Z_r values and the species-specific means of sperm length (for males) and egg volume (for females), respectively, were then examined. These among-species relationships, as well as those of the standardized slopes, were examined using phylogenetic generalized least-squared (PGLS) regressions⁴³ to account for statistical non-independence of data points due to shared ancestry of species, based on the same molecular phylogeny as in the allometric relationships above⁴⁰. Using maximum-likelihood methods, PGLS models estimate the phylogenetic scaling parameter Pagel's λ to evaluate the phylogenetic relationship of the covariance in the residuals⁴³. We used likelihood ratio tests to establish whether the models with the maximum-likelihood value of λ differed from models with values of $\lambda = 0$ or $\lambda = 1$, respectively, with λ close to 0 indicating phylogenetic independence and λ close to 1 indicating a strong phylogenetic association of the traits⁴³.

Quantitative genetic analyses of female preference, male ornament and associated characters. The genetic architecture underlying each trait was evaluated by using the 'animal model' and a resampling approach to estimate the variance components^{44,45}. Means of each of the six families per isoline cross, rather than individual flies, represented our sample size in order to minimize missing data and because, for some traits such as SR and thorax length, we had only one measure per family²⁶. We resampled with replacement among the three family means per isoline cross and block using the SURVEYSELECT procedure in SAS v9.3 (SAS Institute 2011) and calculated their mean for each of 1,000 resampling replicates. For each replicate data set, we then conducted a generalized linear mixed model (procedure GLIMMIX) on these mean values, with block as a fixed effect, paternal and maternal lines and their interactions as random effects, and a multilevel effect defining the nuclear parental contributions. This model is an incomplete diallel with reciprocal but no self crosses^{44,45}: in the diallel analysis it is assumed that the nuclear contributions (N) of the male and females are drawn from the same distribution.

The model decomposed for each replicate the total phenotypic variance into different genetic and residual contributions^{44,45}:

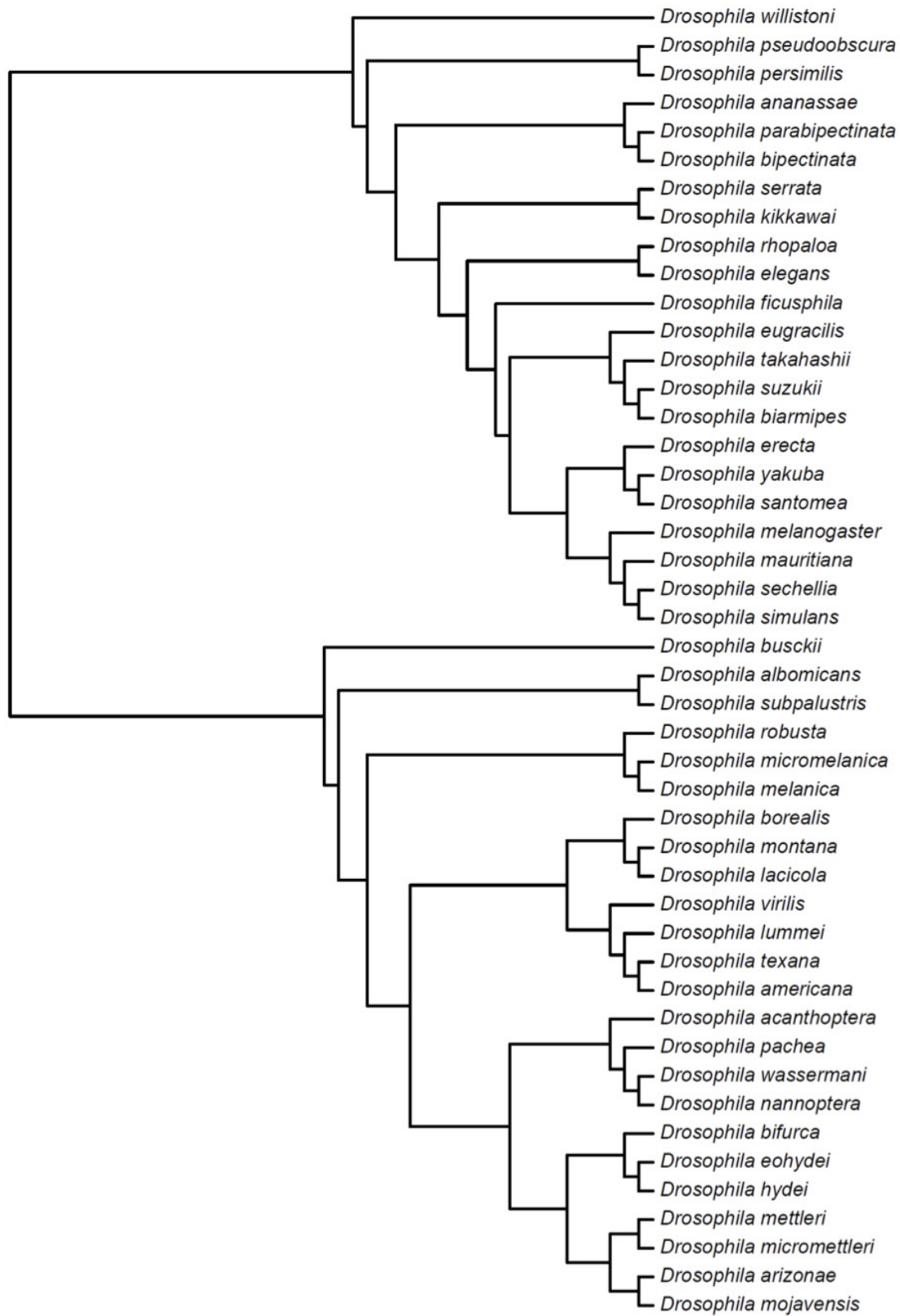
$$Y_{ijk} = \mu + N_i + N_j + T_{ij} + M_j + P_i + K_{ij} + R_{k(ij)}$$

where Y_{ijk} is the trait of the k th replicate cross between isoline i sires and isoline j dams, and μ is the trait mean of the population. N_i and N_j represent the additive contributions by nuclear genes of the respective parental isolines, independent of sex; T_{ij} is the interaction between the haploid nuclear contributions; M_j represents the maternal genetic and environmental effects of isoline j and P_i the paternal genetic and environmental effects of isoline i ; K_{ij} reflects the interaction between maternal and paternal contributions; and $R_{k(ij)}$ is the effect of the k th replicate cross within each combination of dam \times sire lines^{46,47}. Means and standard errors of these variance components across all replicate data sets were then bootstrapped and their statistical significance was determined by testing their z scores (that is, variance component divided by its bootstrapped standard error) against the corresponding significance levels from a standard normal probability table. We used one-tailed significance levels under the a priori constraint that variances are means of squared values, which therefore necessarily have a positive sign.

In the present study, we used only the additive nuclear variance components, σ^2_m , which was necessary to calculate the heritability of, and genetic correlations between, traits of interest. Based on the estimates of the variance components from the diallel analysis, the causal component of the additive nuclear variance, V_A , was estimated as $V_A = 4\sigma^2_m/(1+f)$, where f is the theoretical inbreeding coefficient ($f = 0.96$ based on 15 generations of full-sibling inbreeding³⁷). The additive-by-additive epistatic variance was ignored under the assumption that such higher-order variance is generally very small^{45,48}. Mean values calculated in the above resampling procedure were used to estimate the variances and covariances based on separate univariate analyses of traits x_1 and x_2 , and $x_1 + x_2$, resulting in covariances as $\text{cov}(x_1, x_2) = [\text{var}(x_1 + x_2) - \text{var}(x_1) - \text{var}(x_2)]/2$. We then calculated the corresponding genetic correlations as $r_A = \text{cov}(x_1, x_2)/[\text{var}(x_1) \times \text{var}(x_2)]$ for each of the 1,000 replicates⁴⁹, bootstrapped the genetic correlation coefficient and its standard error, and tested for statistical significance by comparing the z scores to two-tailed significance levels derived from a standard normal distribution⁵⁰.

31. Manier, M. K. *et al.* Resolving mechanisms of competitive fertilization success in *Drosophila melanogaster*. *Science* **328**, 354–357 (2010).
32. Parsons, P. A. & Hosgood, S. M. W. Genetic heterogeneity among the founders of laboratory populations of *Drosophila*. I. Scutellar chaetae. *Genetica* **38**, 328–339 (1968).
33. Falconer, D. S. *Introduction to Quantitative Genetics*. (John Wiley & Sons, 1989).
34. Markow, T. A. & O'Grady, P. *Drosophila: A Guide to Species Identification and Use*. (Academic Press, 2006).
35. Lüpold, S., Tomkins, J. L., Simmons, L. W. & Fitzpatrick, J. L. Female monopolization mediates the relationship between pre- and postcopulatory sexual traits. *Nat. Commun.* **5**, 3184 (2014).
36. Kawano, K. Sexual dimorphism and the making of oversized male characters in beetles (Coleoptera). *Ann. Entomol. Soc. Am.* **99**, 327–341 (2006).
37. Echelle, A. F., Echelle, A. A. & Fitch, H. S. Inter- and intraspecific allometry in a display organ: The dewlap of *Anolis* (Iguanidae) species. *Copeia* **1978**, 245–250 (1978).
38. Simmons, L. W. & Tomkins, J. L. Sexual selection and allometry of earwig forceps. *Evol. Ecol.* **10**, 97–104 (1996).
39. Pitnick, S. & Markow, T. A. Male gametic strategies: sperm size, testes size, and the allocation of ejaculate among successive mates by the sperm-limited fly *Drosophila pacifica* and its relatives. *Am. Nat.* **143**, 785–819 (1994).
40. van der Linde, K., Houle, D., Spicer, G. S. & Stepan, S. J. A supermatrix-based molecular phylogeny of the family Drosophilidae. *Genet. Res.* **92**, 25–38 (2010).
41. Seetharam, A. S. & Stuart, G. W. Whole genome phylogeny for 21 *Drosophila* species using predicted 2b-RAD fragments. *PeerJ* **1**, e226 (2013).
42. Rosenthal, R. *Meta-Analytic Procedures for Social Research*. (Sage, 1991).
43. Freckleton, R. P., Harvey, P. H. & Pagel, M. Phylogenetic analysis and comparative data: a test and review of evidence. *Am. Nat.* **160**, 712–726 (2002).
44. Cockerham, C. C. & Weir, B. S. Quadratic analyses of reciprocal crosses. *Biometrics* **33**, 187–203 (1977).
45. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits*. (Sinauer Associates Inc, 1998).
46. Fry, J. D. In *Genetic Analysis of Complex Traits using SAS* (ed. Saxton, A. M.) 11–34 (SAS Institute Inc., 2004).
47. Bilde, T., Friberg, U., Maklakov, A. A., Fry, J. D. & Arnqvist, G. The genetic architecture of fitness in a seed beetle: assessing the potential for indirect genetic benefits of female choice. *BMC Evol. Biol.* **8**, 295 (2008).
48. Falconer, D. S. & Mackay, T. F. C. *Introduction to Quantitative Genetics*. (Longman, 1996).
49. Crusio, W. E. Bi- and multivariate analyses of diallel crosses: a tool for the genetic dissection of neurobehavioral phenotypes. *Behav. Genet.* **23**, 59–67 (1993).
50. Juenger, T. & Bergelson, J. The evolution of compensation to herbivory in scarlet gilia, *Ipomopsis aggregata*: herbivore-imposed natural selection and the quantitative genetics of tolerance. *Evolution* **54**, 764–777 (2000).
51. Madge, S. & McGowan, P. *Pheasants, Partridges, and Grouse: A Guide to the Pheasants, Partridges, Quails, Grouse, Guineafowl, Buttonquails, and Sandgrouse of the World*. (Princeton University Press, 2002).
52. Eo, S. H., Bininda-Emonds, O. R. P. & Carroll, J. P. A phylogenetic supertree of the fowls (Galloanserae, Aves). *Zool. Scr.* **38**, 465–481 (2009).
53. Lemaître, J. F., Vanpé, C., Plard, F. & Gaillard, J. M. The allometry between secondary sexual traits and body size is nonlinear among cervids. *Biol. Lett.* **10**, 20130869 (2014).
54. Moen, R. A., Pastor, J. & Cohen, Y. Antler growth and extinction of Irish elk. *Evol. Ecol. Res.* **1**, 235–249 (1999).
55. Gould, S. J. The origin and function of 'bizarre' structures: Antler size and skull size in the 'Irish Elk', *Megaloceros giganteus*. *Evolution* **28**, 191–220 (1974).
56. Plard, F., Bonenfant, C. & Gaillard, J.-M. Revisiting the allometry of antlers among deer species: male-male sexual competition as a driver. *Oikos* **120**, 601–606 (2011).
57. Bro-Jørgensen, J. The intensity of sexual selection predicts weapon size in male bovids. *Evolution* **61**, 1316–1326 (2007).
58. Lüpold, S., Simmons, L. W., Tomkins, J. L. & Fitzpatrick, J. L. No evidence for a trade-off between sperm length and male premating weaponry. *J. Evol. Biol.* **28**, 2187–2195 (2015).
59. Agnarsson, I. & May-Collado, L. J. The phylogeny of Cetartiodactyla: the importance of dense taxon sampling, missing data, and the remarkable promise of cytochrome b to provide reliable species-level phylogenies. *Mol. Phylogenet. Evol.* **48**, 964–985 (2008).
60. Arnold, C., Matthews, L. J. & Nunn, C. L. The 10kTrees website: a new online resource for primate phylogeny. *Evol. Anthropol.* **19**, 114–118 (2010).
61. Bergmann, P. J. & Berk, C. P. The evolution of positive allometry of weaponry in horned lizards (Phrynosoma). *Evol. Biol.* **39**, 311–323 (2012).
62. Rowland, J. M. & Miller, K. B. Phylogeny and systematics of the giant rhinoceros beetles (Scarabaeidae: Dynastini). *Insecta Mundi* **0263**, 1–15 (2012).
63. Simmons, L. W. & Emlen, D. J. Evolutionary trade-off between weapons and testes. *Proc. Natl Acad. Sci. USA* **103**, 16346–16351 (2006).
64. Baker, R. H. & Wilkinson, G. S. Phylogenetic analysis of sexual dimorphism and eye-span allometry in stalk-eyed flies (Diopsidae). *Evolution* **55**, 1373–1385 (2001).
65. Knell, R. J., Pomfret, J. C. & Tomkins, J. L. The limits of elaboration: curved allometries reveal the constraints on mandible size in stag beetles. *Proc. R. Soc. Lond. B* **271**, 523–528 (2004).

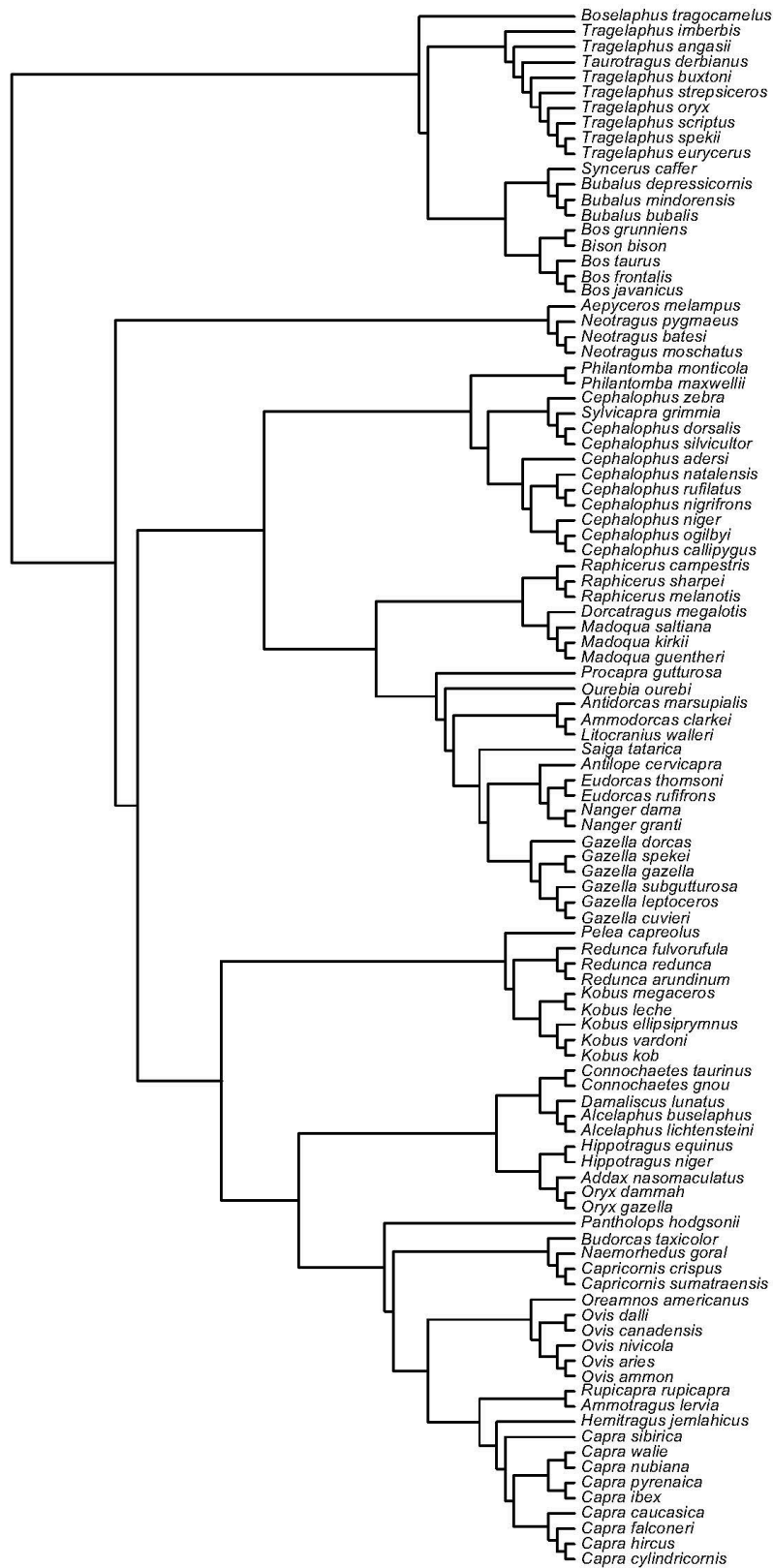
66. Sharma, P. P., Buenavente, P. A. C., Clouse, R. M., Diesmos, A. C. & Giribet, G. Forgotten gods: Zalmoxidae of the Philippines and Borneo. *Zootaxa* **3280**, 29–55 (2012).
67. Roewer, C. F. & Weitere Weberknechte I. (1. Ergänzung der: 'Weberknechte der Erde,' 1923). *Abhandlungen des Naturwissenschaftlichen Vereins zu Bremen* **26**, 261–402 (1927).
68. Forster, R. R. Further Australian harvestmen (Arachnida: Opiliones). *Aust. J. Zool.* **3**, 354–411 (1955).
69. Sharma, P. P. New Australasian Zalmoxidae (Opiliones: Laniatores) and a new case of male polymorphism in Opiliones. *Zootaxa* **3236**, 1–35 (2012).
70. Roewer, C. F. Die Weberknechte der Erde. *Systematische Bearbeitung der bisher bekannten Opiliones*. (Verlag von Oustav fiseher, 1923).
71. Sharma, P. P. & Giribet, G. Out of the Neotropics: Late Cretaceous colonization of Australasia by American arthropods. *Proc. R. Soc. Lond. B* **279**, 3501–3509 (2012).



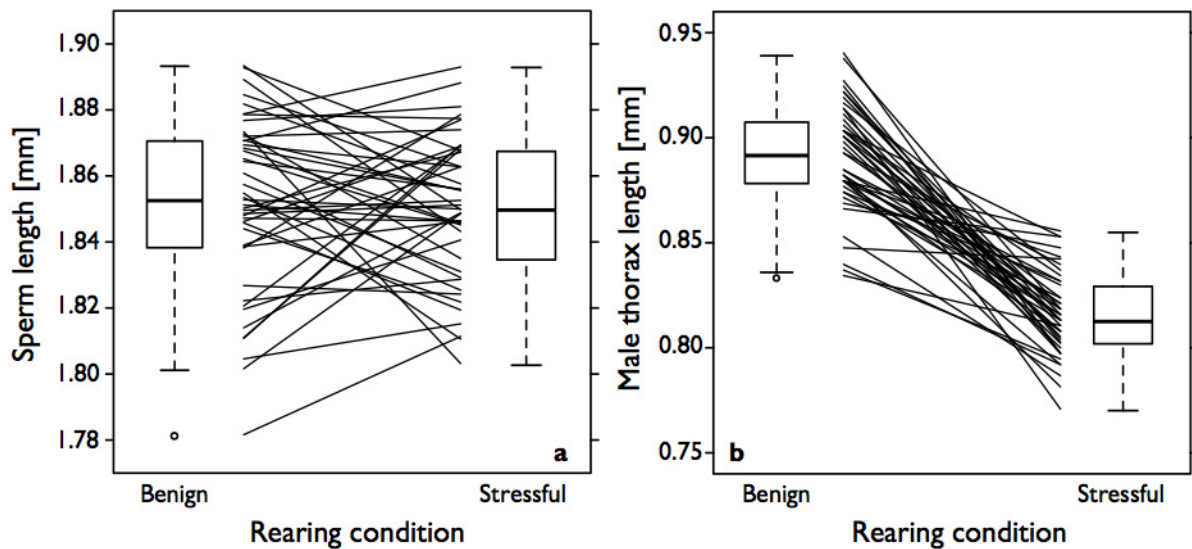
Extended Data Figure 1 | Phylogeny for the *Drosophila* comparative analyses of gamete allometry. Molecular phylogeny of the 46 species based on ref. 40, with species added based on refs. 29 and 41. Owing to a lack of information on branch lengths, equal branch lengths were used.



Extended Data Figure 2 | Phylogeny of the Phasianinae. Tree topology of the Phasianinae in Supplementary Table 1 based on the molecular phylogeny of ref. 52. Owing to a lack of information on branch lengths, equal branch lengths were used.

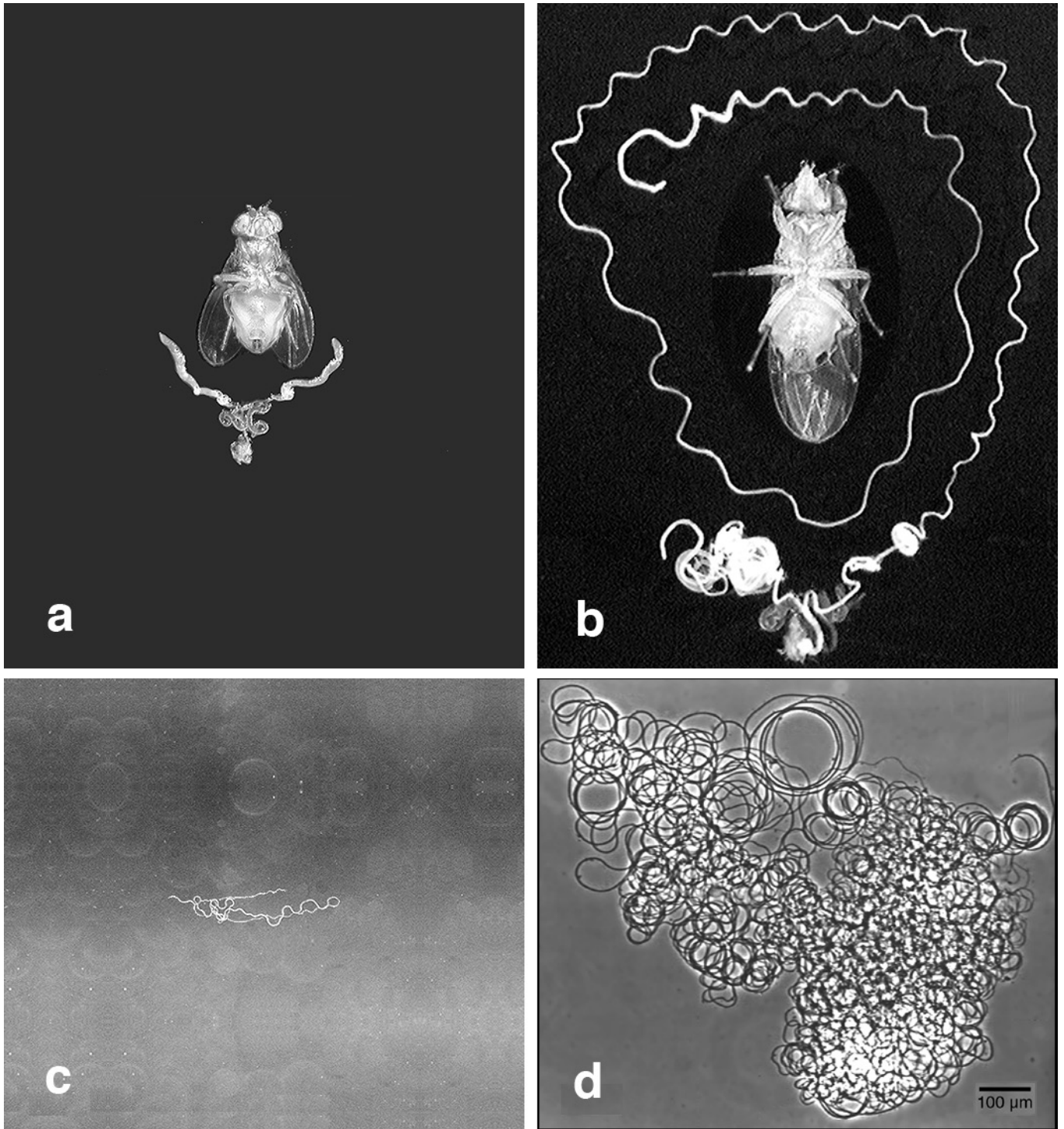


Extended Data Figure 3 | Phylogeny of the Bovidae. Tree topology of the Bovidae in Supplementary Table 2 based on the molecular phylogenies of the 10kTrees Project⁶⁰ and ref. 59. Equal branch lengths were used because of combining different trees.

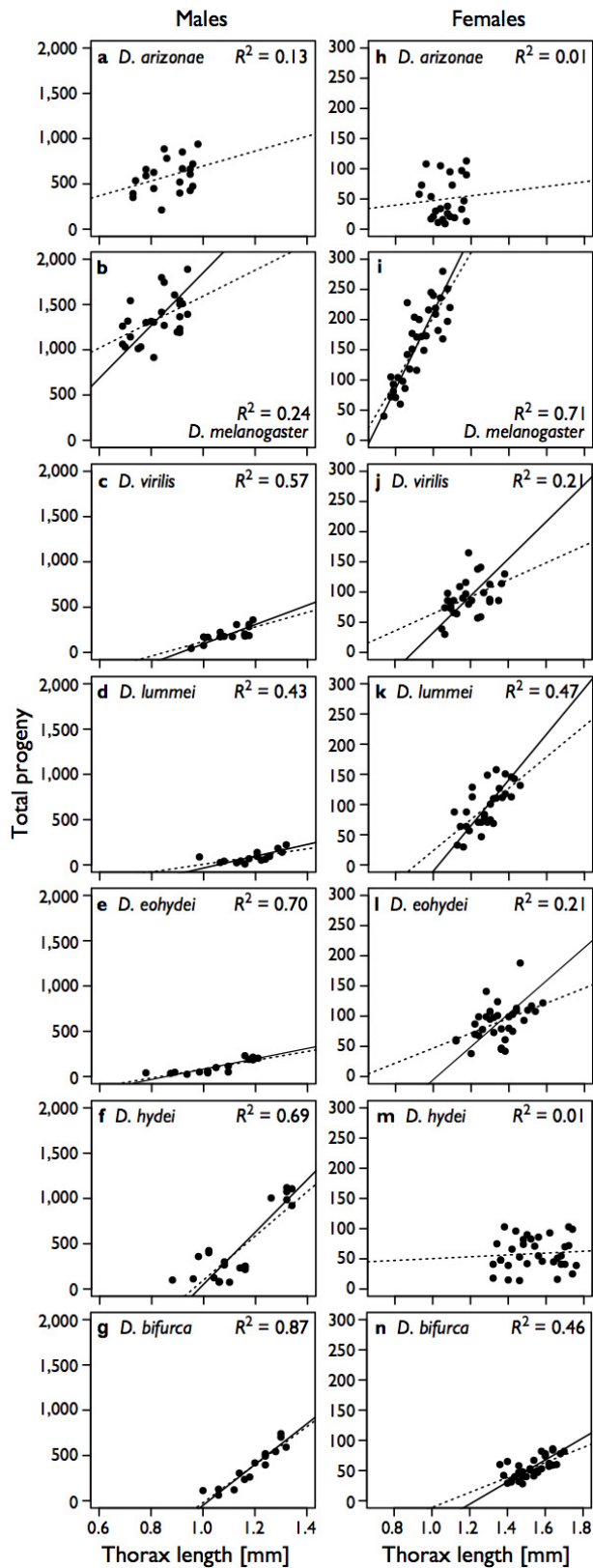


Extended Data Figure 4 | Lacking condition dependence of sperm length. a, b, Comparison of sperm length (a) and male thorax length (b) between flies reared under benign and moderately stressful conditions. Each line connects the means of a nuclear genotype ($n = 45$), based on measurements of the same five males in a and b, and the box plots reflect the between-genotype variation for each treatment. On average, sperm length did not differ between the benign (mean \pm s.d. = 1.853 ± 0.019 mm)

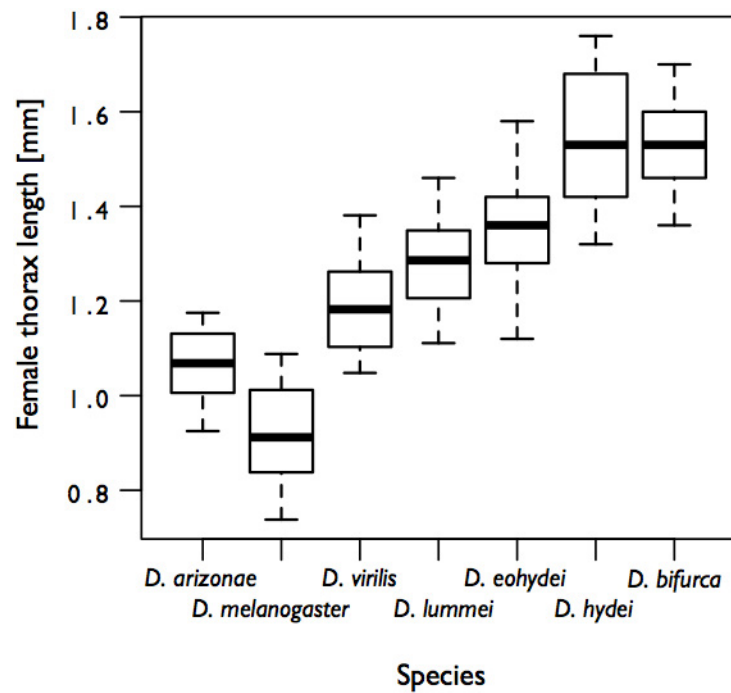
and moderately stressful treatments (1.851 ± 0.021 mm; linear mixed-effects model controlling for genetic background: $t = -0.57$, $P = 0.58$), thereby reflecting no condition dependence. By contrast, all males reared under stressful conditions were smaller (thorax length: 0.816 ± 0.019 mm versus 0.892 ± 0.026 mm; $t = -17.08$, $P < 0.0001$), thus being strongly condition-dependent and highlighting the relatively higher cost of sperm length for low-quality males.



Extended Data Figure 5 | Variation in investment per sperm and in spermatogenesis. a–d, Intact male fly above his reproductive tract (a, b) and a single spermatozoon (c, d) for *Drosophila arizonae* (a, c) and *D. bifurca* (b, d). Top panels and bottom panels depict equal magnification, respectively. All photos by S.P.



Extended Data Figure 6 | Condition dependence of male and female reproductive potential in seven *Drosophila* species. a–n, Intraspecific relationships between reproductive potential and body size as a proxy of condition for males (a–g) and females (h–n) of seven *Drosophila* species. Species are ordered from shortest (top) to longest (bottom) sperm. Dotted lines represent ordinary least-squares slopes and, where these regressions were statistically significant, solid lines indicate RMA slopes. For detailed statistics see Extended Data Table 3.



Extended Data Figure 7 | Comparison of intraspecific variation in female thorax length. Box plot reflecting the greater intraspecific standard deviation in female thorax length in *D. hydei* compared to the remaining species (Bartlett's test of homogeneity of variances: $K^2 = 12.67$, $P = 0.05$).

Extended Data Table 1 | Statistics of evolutionary allometries in different taxa

Taxon	Sexual trait	Size trait	N	Slope	P	λ	Source
<i>Drosophila</i>	Sperm length	Thorax length	46	5.52	<0.0001	1.00	This study
Phasianinae	Spur length	Body mass	42	2.26*	<0.0001	1.00	35,51,52 [†]
Phasianinae	Spur length	Tarsus length	40	2.00	<0.0001	0.98	35,51,52 [†]
Phasianinae	Tail length	Body mass	51	3.75*	<0.0001	1.00	35,51,52 [†]
Phasianinae	Tail length	Tarsus length	54	3.20	<0.0001	0.98	35,51,52 [†]
Cervidae	Antler length	Body mass	31	2.98*	<0.0001	0.62	35,53,54 [‡]
Cervidae	Antler mass	Body mass	21	1.73*	<0.0001	0.11	35,53,54 [‡]
Cervidae	Antler length	Shoulder height	20	1.85			55
Cervidae	Antler length	Body mass	31	1.80*			56 [§]
Bovidae	Horn length	Body mass	102	2.24*	<0.0001	0.93	57–60
Bovidae	Horn length	Shoulder height	76	2.22			55 [¶]
<i>Phrynosoma</i> lizards	Horn length	Snout-vent length	14	2.88			61
Dynastini rhinoceros beetles	Horn length	Body length	12	1.96	0.02	1.00	36,62 [#]
Onthophagini dung beetles	Horn length	Pronotum width	22	1.86	0.002	<0.001	63
<i>Anolis</i> lizards	Dewlap area	Snout-vent length	22	1.70*			37
Diopsidae	Eye span	Body length	30	2.74	0.0004	0.58	64
<i>Cyclommatus</i> stag beetles	Mandible length	Body length	10	1.87			14
<i>Lucanus</i> stag beetles	Mandible length	Elythra length	17	2.06			65
<i>Neolucanus</i> stag beetles	Mandible length	Body length	11	1.37			14
Dermaptera	Forcep length	Pronotum width	42	1.55			38
<i>Zalmoxis</i> harvestmen	Hind-leg length	Body length	16	1.55	0.023	<0.001	66–71 ^{**}
<i>Zalmoxis</i> harvestmen	Hind-leg length	Prosoma width	11	2.47	0.002	<0.001	66,69,71 ^{**}

Interspecific allometric slopes between different sexually selected traits and body-size indices are listed along with the number of species (*N*), the *P* value of the regression analysis against a slope of 1, and the phylogenetic scaling factor λ . Where *P* and λ values are present, slopes were calculated using a phylogenetic RMA analysis based on the data and phylogenies from the cited sources. All other slopes are taken directly from the corresponding sources. References 14,35–38,51–71 are cited in this table.

*For direct comparison between slopes, these analyses were adjusted to have an isometric slope of 1 by cube-rooting mass variables or square-rooting area variables.

†For species and phylogeny see Supplementary Information Table 2 and Fig. 2.

‡Despite reports on allometric slopes in ref. 53, these slopes were reanalysed using phylogenetic RMA regressions and with the Irish elk (*Megaceros giganteus*) included⁵⁴.

§Does not include the Irish elk (*M. giganteus*).

||Only slope of ordinary least-squares regression were reported, and no data for reanalysis.

¶For species and phylogeny see Supplementary Information Table 3 and Fig. 3.

#Some species with data were not listed in the phylogeny⁶², but they could assume the position of their single congeneric representative in the phylogeny. Only the relative position of the three *Chalcosoma* species was unclear, and they were thus combined in the same node.

**For species and phylogeny see Supplementary Information Table 4.

Extended Data Table 2 | Comparative data set of *Drosophila* gamete and body size

Species	Male thorax length [mm]	Sperm length [mm]	Female thorax length [mm]	Egg volume [mm ³ × 1,000]	Ovariolo number
<i>D. acanthoptera</i>	1.13	5.83	1.15	7.10	41.88
<i>D. albomicans</i>	1.25	5.35	1.40	6.31	
<i>D. americana</i>	1.38	5.22	1.29	8.20	31.60
<i>D. ananassae</i>	0.93	2.16	1.03	5.18	23.00
<i>D. arizonae</i>	0.95	1.52	1.05	5.40	34.60
<i>D. biarmipes</i>	0.93	1.91	1.00	6.53	
<i>D. bifurca</i>	1.60	58.29	1.53	8.20	51.53
<i>D. bipectinata</i>	0.83	1.75	0.92	5.30	25.00
<i>D. borealis</i>	1.28	7.54	1.29	9.00	30.78
<i>D. busckii</i>	0.82	1.18	0.98	3.40	52.86
<i>D. elegans</i>	0.92	2.22	0.96	8.78	
<i>D. eohydei</i>	1.39	18.11	1.28	7.10	39.22
<i>D. erecta</i>	0.71	1.15	0.93	5.17	
<i>D. eugracilis</i>	1.07	2.10	1.19	10.11	
<i>D. ficusphila</i>	1.09	1.80	1.21	5.83	13.33
<i>D. hydei</i>	1.33	23.32	1.43	8.00	51.75
<i>D. kikkawai</i>	0.87	2.87	0.96	3.85	
<i>D. laticola</i>	1.21	2.52	1.22	6.60	28.67
<i>D. lummei</i>	1.34	7.79	1.42	9.70	36.00
<i>D. mauritiana</i>	0.86	0.98	0.95	5.87	
<i>D. melanica</i>	1.12	4.93	1.33	10.50	42.20
<i>D. melanogaster</i>	0.88	1.85	0.98	8.30	33.14
<i>D. mettleri</i>	1.17	2.79	1.20	5.00	44.75
<i>D. micromelanica</i>	0.98	1.41	1.16	8.70	25.57
<i>D. micromettleri</i>	0.95	2.22	1.14	7.00	36.29
<i>D. mojavensis</i>	0.89	1.90	1.03	5.40	33.17
<i>D. montana</i>	1.41	3.34	1.32	9.80	28.80
<i>D. nanoptera</i>	0.99	15.74	1.06	7.30	37.60
<i>D. pachea</i>	1.02	16.53	0.99	6.00	28.17
<i>D. parabiplectinata</i>	0.84	1.93	0.99	5.15	
<i>D. persimilis</i>	0.93	0.32	1.06	8.60	36.00
<i>D. pseudoobscura</i>	1.01	0.36	1.09	6.20	45.42
<i>D. rhopaloea</i>	0.97	2.43	1.08	12.94	
<i>D. robusta</i>	1.44	6.63	1.47	7.10	41.25
<i>D. santomea</i>	0.92	1.11	1.02	7.07	
<i>D. sechellia</i>	0.81	1.74	0.91	8.18	
<i>D. serrata</i>	0.86	3.63	1.00	4.02	
<i>D. simulans</i>	0.87	1.10	0.89	7.40	36.83
<i>D. subpalustris</i>	1.23	5.96	1.35	11.30	26.20
<i>D. suzukii</i>	1.10	1.67	1.28	6.93	
<i>D. takahashii</i>	0.90	1.87	1.06	6.13	
<i>D. texana</i>	1.29	5.08	1.27	6.80	38.00
<i>D. virilis</i>	1.27	5.70	1.25	8.50	41.21
<i>D. wassermani</i>	1.07	4.52	1.15	6.70	33.88
<i>D. willistoni</i>	0.95	6.62	0.90	7.30	22.60
<i>D. yakuba</i>	0.81	1.75	0.90	5.63	

Species means of male and female traits used in the comparative analyses. Data were taken from references 21, 29 and 34, except for egg data where ovariolo numbers are missing (measured in current study). Species used in the comparative analyses of the sex-specific condition dependence of reproductive potential are indicated by bold typeface.

Extended Data Table 3 | Intraspecific analyses of condition dependence of reproductive potential

Species	N	r	Slope	t	P	Weighted Z_r
Males						
<i>D. arizonae</i>	20	0.364	0.364	1.659	0.1144	6.873
<i>D. melanogaster</i>	26	0.492	0.492	2.766	0.0107	12.919
<i>D. virilis</i>	16	0.753	0.753	4.285	0.0008	13.728
<i>D. lummei</i>	15	0.653	0.653	3.108	0.0083	10.144
<i>D. eohydei</i>	16	0.839	0.839	5.771	<0.0001	17.054
<i>D. hydei</i>	23	0.832	0.832	6.875	<0.0001	25.093
<i>D. bifurca</i>	15	0.933	0.933	9.362	<0.0001	21.874
Females						
<i>D. arizonae</i>	24	0.084	0.084	0.396	0.6962	1.852
<i>D. melanogaster</i>	34	0.842	0.842	8.812	<0.0001	39.241
<i>D. virilis</i>	28	0.459	0.459	2.634	0.0140	12.894
<i>D. lummei</i>	30	0.685	0.685	4.968	<0.0001	23.451
<i>D. eohydei</i>	33	0.454	0.454	2.838	0.0079	15.189
<i>D. hydei</i>	34	0.081	0.081	0.459	0.6495	2.591
<i>D. bifurca</i>	34	0.680	0.680	5.242	<0.0001	26.514

Statistical results of the intraspecific analyses of the male or female reproductive potential against the corresponding body size as a proxy of physical condition. Slopes are standardized for each species (that is, all variables centred around 0 and divided by corresponding standard deviation).

Genome-wide association study identifies 74 loci associated with educational attainment

A list of authors and their affiliations appears in the online version of the paper.

Educational attainment is strongly influenced by social and other environmental factors, but genetic factors are estimated to account for at least 20% of the variation across individuals¹. Here we report the results of a genome-wide association study (GWAS) for educational attainment that extends our earlier discovery sample^{1,2} of 101,069 individuals to 293,723 individuals, and a replication study in an independent sample of 111,349 individuals from the UK Biobank. We identify 74 genome-wide significant loci associated with the number of years of schooling completed. Single-nucleotide polymorphisms associated with educational attainment are disproportionately found in genomic regions regulating gene expression in the fetal brain. Candidate genes are preferentially expressed in neural tissue, especially during the prenatal period, and enriched for biological pathways involved in neural development. Our findings demonstrate that, even for a behavioural phenotype that is mostly environmentally determined, a well-powered GWAS identifies replicable associated genetic variants that suggest biologically relevant pathways. Because educational attainment is measured in large numbers of individuals, it will continue to be useful as a proxy phenotype in efforts to characterize the genetic influences of related phenotypes, including cognition and neuropsychiatric diseases.

Educational attainment is measured in all main analyses as the number of years of schooling completed (EduYears, $n = 293,723$, mean = 14.3, s.d. = 3.6; Supplementary Information sections 1.1–1.2). All GWAS were performed at the cohort level in samples restricted to individuals of European descent whose educational attainment was assessed at or above age 30. A uniform set of quality-control procedures was applied to the cohort-level summary statistics. In our GWAS meta-analysis of ~ 9.3 million SNPs from the 1000 Genomes Project, we used sample-size weighting and applied a single round of genomic control at the cohort level.

Our meta-analysis identified 74 approximately independent genome-wide significant loci. For each locus, we define the ‘lead SNP’ as the SNP in the genomic region that has the smallest P value (Supplementary Information section 1.6.1). Figure 1 shows a Manhattan plot with the lead SNPs highlighted. This includes the three SNPs that reached genome-wide significance in the discovery stage of our previous GWAS meta-analysis of educational attainment¹. The quantile–quantile (Q–Q) plot of the meta-analysis (Extended Data Fig. 1) exhibits inflation ($\lambda_{GC} = 1.28$), as expected under polygenicity³.

Extended Data Fig. 2 shows the estimated effect sizes of the lead SNPs. The estimates range from 0.014 to 0.048 standard deviations per allele (2.7 to 9.0 weeks of schooling), with incremental R^2 in the range 0.01% to 0.035%.

To quantify the amount of population stratification in the GWAS estimates that remains even after the stringent controls used by the cohorts (Supplementary Information section 1.4), we used linkage-disequilibrium (LD) score regression⁴. The regression results indicate that $\sim 8\%$ of the observed inflation in the mean χ^2 is due to bias rather than polygenic signal (Extended Data Fig. 3a), suggesting that stratification effects are small in magnitude. We also found evidence for polygenic association signal in several within-family analyses, although these are not powered for individual SNP association testing (Supplementary Information section 2 and Extended Data Fig. 3b).

To further test the robustness of our findings, we examined the within-sample and out-of-sample replicability of SNPs reaching genome-wide significance (Supplementary Information sections 1.7–1.8). We found that SNPs identified in the previous educational attainment meta-analysis replicated in the new cohorts included here, and conversely, that SNPs reaching genome-wide significance in the new cohorts replicated in the old cohorts. For the out-of-sample replication analyses of our 74 lead SNPs, we used the interim release of the UK Biobank⁵ (UKB) ($n = 111,349$). As shown in Extended Data Fig. 4,

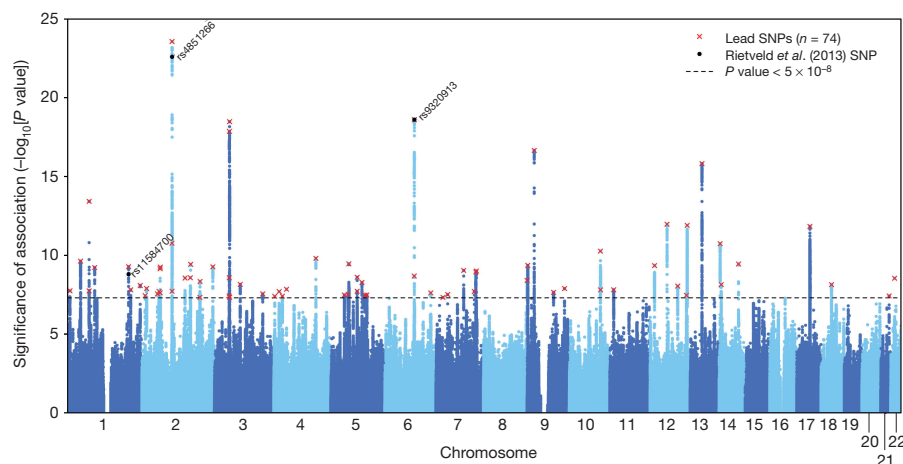


Figure 1 | Manhattan plot for EduYears associations ($n = 293,723$). The x axis is chromosomal position, and the y axis is the significance on a $-\log_{10}$ scale (two-tailed test). The black dashed line shows the genome-

wide significance level (5×10^{-8}). The red crosses are the 74 approximately independent genome-wide significant associations (lead SNPs). The black dots labelled with rs numbers are the three SNPs identified in ref. 1.

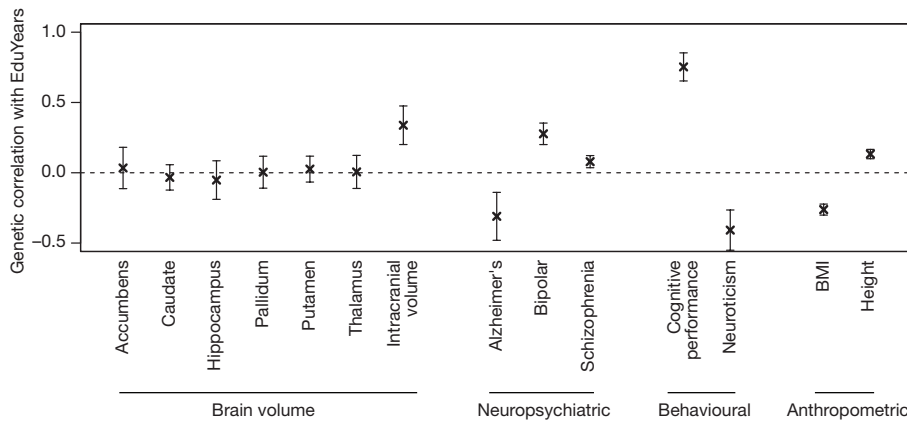


Figure 2 | Genetic correlations between EduYears and other traits. Results from bivariate LD score regressions⁹: estimates of genetic correlation with brain volume, neuropsychiatric, behavioural, and anthropometric phenotypes using published GWAS summary statistics. The error bars show the 95% confidence intervals (CI).

72 out of the 74 lead SNPs have a consistent sign ($P = 1.47 \times 10^{-19}$), 52 are significant at the 5% level ($P = 2.68 \times 10^{-50}$), and 7 reach genome-wide significance in the UK Biobank data set ($P = 1.41 \times 10^{-42}$). For comparison, the corresponding expected numbers, assuming each SNP's true effect size is its estimated effect adjusted for the winner's curse, are 71.4, 40.3, and 0.6. (Supplementary Information section 1.8.2). We also find out-of-sample replicability of our overall GWAS results: the genetic correlation between EduYears in our meta-analysis sample and in the UKB data is 0.95 (s.e. = 0.021; Supplementary Table 1.14).

It is known that educational attainment, cognitive performance, and many neuropsychiatric phenotypes are phenotypically correlated, and several studies of twins find that the phenotypic correlations partly reflect genetic overlap^{6–8} (Supplementary Information section 3.3.4). Here we investigate genetic correlation using our GWAS results for EduYears and published GWAS results for 14 other phenotypes, using bivariate LD score regression⁹ (Supplementary Information section 3). First, we estimated genetic correlations with EduYears. As shown in Fig. 2, based on overall summary statistics for associated variants, we find genetic covariance between increased educational attainment and increased cognitive performance ($P = 9.9 \times 10^{-50}$), increased intracranial volume ($P = 1.2 \times 10^{-6}$), increased risk of bipolar disorder ($P = 7 \times 10^{-13}$), decreased risk of Alzheimer's ($P = 4 \times 10^{-4}$), and lower neuroticism ($P = 2.8 \times 10^{-8}$). We also found positive, statistically significant, but very small, genetic correlations with height ($P = 5.2 \times 10^{-15}$) and risk of schizophrenia ($P = 3.2 \times 10^{-4}$).

Second, we examined whether our 74 lead SNPs are jointly associated with each phenotype (Extended Data Fig. 5 and Supplementary Information section 3.3.1). We reject the null hypothesis of no enrichment at $P < 0.05$ for 10 of the 14 phenotypes (all the exceptions are subcortical brain structures).

Third, for each phenotype, we tested (in the published GWAS results) each of our 74 lead SNPs (or its proxy) for association at a significance threshold of 0.05/74. We found a total of 25 SNPs meeting this threshold for any of these phenotypes, but only one reaching genome-wide significance. While these results provide suggestive evidence that some of these SNPs may be associated with other phenotypes, further testing of these associations in independent cohorts is required (Supplementary Tables 3.2–3.4, Extended Data Fig. 6).

To consider potential biological pathways, we first tested whether SNPs in particular regions of the genome are implicated by our GWAS results. Unlike what has been found for other phenotypes, SNPs in regions that are DNase I hypersensitive in the fetal brain are more likely to be associated with EduYears by a factor of ~ 5 (95% confidence interval 2.89–7.07; Extended Data Fig. 7). Moreover, the 15% of SNPs residing in regions associated with histones marked in the central nervous

system (CNS) explain 44% of the heritable variation (Extended Data Fig. 8a and Supplementary Table 4.4.2). This enrichment factor of ~ 3 for CNS ($P = 2.48 \times 10^{-16}$) is greater than that of any of the other nine tissue categories in this analysis.

Given that our findings disproportionately implicate SNPs in regions regulating brain-specific gene expression, we examined whether genes located near EduYears-associated SNPs show elevated expression in neural tissue. We tested this hypothesis using data on mRNA transcript levels in the 37 adult tissues assayed by the Genotype-Tissue Expression Project (GTEx)¹⁰. Remarkably, the 13 GTEx tissues that are components of the CNS—and only those 13 tissues—show significantly elevated expression levels of genes near EduYears-associated SNPs (false discovery rate < 0.05 ; Extended Data Fig. 8b and Supplementary Table 4.5.2).

To investigate possible functions of the candidate genes from the GWAS-implicated loci, we examined the extent of their overlap with groups of genes ('gene sets') whose products are known or predicted to participate in a common biological process¹¹. We found 283 gene sets significantly enriched by the candidate genes identified in our GWAS (false discovery rate < 0.05 ; Supplementary Table 4.5.1). To facilitate interpretation, we used a standard procedure¹¹ to group the 283 gene sets into 'clusters' defined by degree of gene overlap. The resulting 34 clusters, shown in Fig. 3, paint a coherent picture, with many clusters corresponding to stages of neural development: the proliferation of neural progenitor cells and their specialization (the cluster npBAF complex), the migration of new neurons to the different layers of the cortex (forebrain development, abnormal cerebral cortex morphology), the projection of axons from neurons to their signalling targets (axonogenesis, signalling by Robo receptor), the sprouting of dendrites and their spines (dendrite, dendritic spine organization), and neuronal signalling and synaptic plasticity throughout the lifespan (voltage-gated calcium channel complex, synapse part, synapse organization).

Many of our results implicate candidate genes and biological pathways that are active during distinct stages of prenatal brain development. To directly examine how the expression levels of candidate genes identified in our GWAS vary over the course of development, we used gene expression data from the BrainSpan Developmental Transcriptome¹². As shown in Extended Data Fig. 9, these candidate genes exhibit above-baseline expression in the brain throughout life but especially higher expression levels in the brain during prenatal development (1.36 times higher prenatally than postnatally, $P = 6.02 \times 10^{-8}$).

A summary overview of some promising candidate genes for follow-up work is provided in Table 1.

We constructed polygenic scores¹³ to assess the joint predictive power afforded by the GWAS results (Supplementary Information section 5.2). Across our two holdout samples, the mean predictive

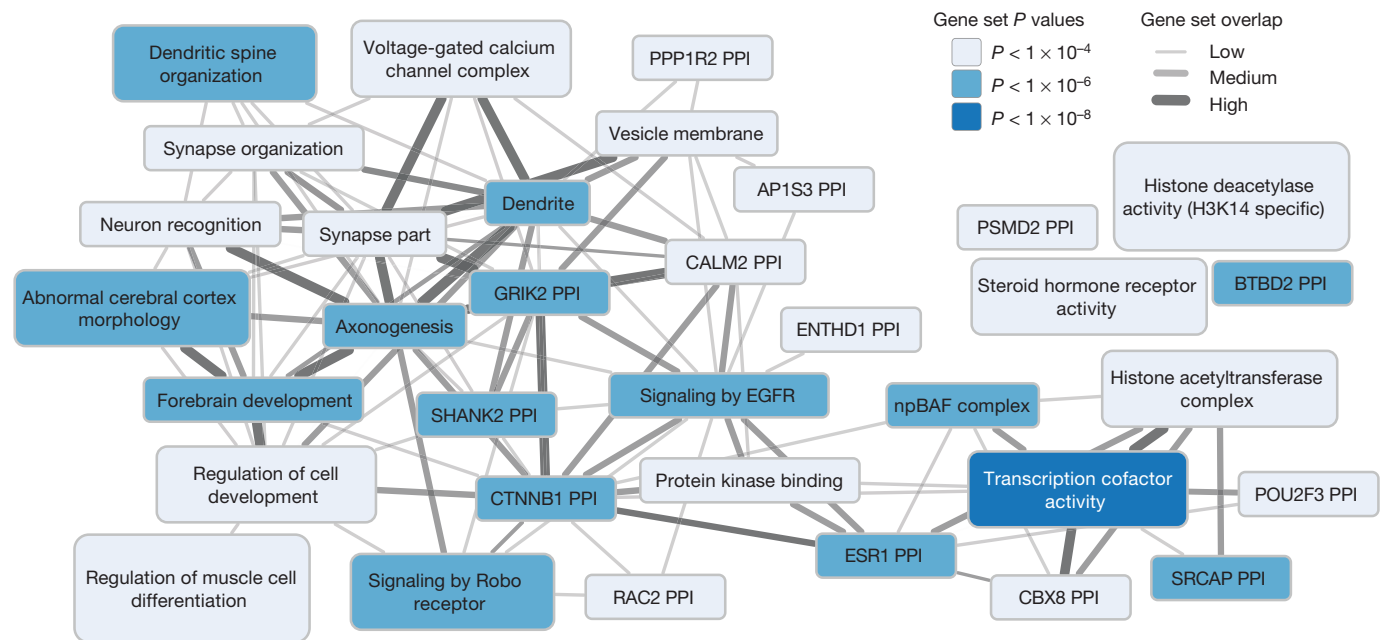


Figure 3 | Overview of biological annotation. Thirty-four clusters of significantly enriched gene sets. Each cluster is named after one of its member gene sets. The colour represents the permutation P value of the member set exhibiting the most statistically significant enrichment. Overlap between pairs of clusters is represented by an edge. Edge width

represents the Pearson correlation ρ between the two vectors of gene membership scores ($\rho < 0.3$, no edge; $0.3 \leq \rho < 0.5$, thin edge; $0.5 \leq \rho < 0.7$, intermediate edge; $\rho \geq 0.7$, thick edge), where each cluster's vector is the vector for the gene set after which the cluster is named.

power of a polygenic score constructed from all measured SNPs is 3.2% ($P = 1.18 \times 10^{-39}$; Supplementary Table 5.2 and Supplementary Information section 5).

Studies of genetic analyses of behavioural phenotypes have been prone to misinterpretation, such as characterizing identified associated variants as ‘genes for education’. Such characterization is not correct for many reasons: educational attainment is primarily determined by environmental factors, the explanatory power of the individual SNPs is small, the candidate genes may not be causal, and the genetic associations with educational attainment are mediated by multiple intermediate phenotypes¹⁴. To illustrate this last point, we studied mediation of the association between the all-SNPs polygenic score and EduYears in two of our cohorts. We found that cognitive performance can statistically account for 23–42% of the

association ($P < 0.001$) and the personality trait ‘openness to experience’ for approximately 7% ($P < 0.001$; Supplementary Information section 6).

It would also be a mistake to infer from our findings that the genetic effects operate independently of environmental factors. Indeed, a recent meta-analysis of twin studies found that genetic influences on educational attainment are heterogeneous across countries and birth cohorts¹⁵. We conducted exploratory analyses in the Swedish Twin Registry to illustrate how environmental factors may amplify or dampen the impact of genetic influences (Supplementary Information section 7). We found that the predictive power of the all-SNPs polygenic score is heterogeneous by birth cohort, with smaller explanatory power in younger cohorts (Extended Data Fig. 10; see also Supplementary Information section 7.4 for discussion of the contrast between these

Table 1 | Selected candidate genes implicated by bioinformatics analyses

Gene	SNP	Syndromic	Score	Top-ranking gene sets
<i>TBR1</i>	rs4500960	ID, ASD	6	Developmental biology, decreased brain size, abnormal cerebral cortex morphology
<i>MEF2C</i>	rs7277187	ID, ASD	5	ErbB signalling pathway, abnormal sternum ossification, regulation of muscle cell differentiation
<i>ZSWIM6</i>	rs61160187	–	5	Transcription factor binding, negative regulation of signal transduction, PI3K events in ErbB4 signalling
<i>BCL11A</i>	rs2457660	ASD	5	Dendritic spine organization, abnormal hippocampal mossy fibre morphology, SWI/SNF-type complex
<i>CELSR3</i>	rs11712056	SCZ	5	Dendrite morphogenesis, dendrite development, abnormal hippocampal mossy fibre morphology
<i>MAPT</i>	rs192818565	ID	5	Dendrite morphogenesis, abnormal hippocampal mossy fibre morphology, abnormal axon guidance
<i>SBNO1</i>	rs7306755	SCZ	5	Protein serine/threonine phosphatase complex
<i>NBAS</i>	rs12987662	–	5	–
<i>NBEA</i>	rs9544418	SCZ	4	Developmental biology, signalling by Robo receptor, dendritic shaft
<i>SMARCA2</i>	rs1871109	ID	4	–
<i>MAP4</i>	rs11712056	ASD	4	Developmental biology, signalling by Robo receptor, SWI/SNF-type complex
<i>LINC00461</i>	rs10061788	–	4	Decreased brain size, abnormal cerebral cortex morphology, abnormal hippocampal mossy fibre morphology
<i>POU3F2</i>	rs9320913	–	4	Dendrite morphogenesis, developmental biology, decreased brain size
<i>RAD54L2</i>	rs11712056	SCZ	4	Decreased brain size, SWI/SNF-type complex, nBAF complex
<i>PLK2</i>	rs2964197	–	4	Negative regulation of signal transduction, PI3K events in ErbB4 signalling

Fifteen candidate genes implicated most consistently across various analyses. To assemble this list, each gene in a DEPICT-defined locus (Supplementary Information section 4.5) was assigned a score equal to the number of criteria it satisfies out of ten (see Supplementary Table 4.1 for details). The DEPICT prioritization P value was used as the tiebreaker. SNP, the SNP in the gene's locus with the lowest P value in the EduYears meta-analysis. Syndromic, which, if any, of three neuropsychiatric disorders have been linked to *de novo* mutations in the gene (Supplementary Information section 4.6). Top-ranking gene sets, DEPICT reconstituted gene sets of which the gene is a top-20 member (Supplementary Table 4.5.1). The three most significant gene sets are shown if more than three are available. ID, intellectual disability; ASD, autism spectrum disorder; SCZ, schizophrenia; ErbB, erythroblastosis oncogene B; PI3K, phosphatidylinositol-4,5-bisphosphate 3-kinase; SWI/SNF, SWI/sucrose non-fermentable; nBAF, neuronal BRG1- or HRBM-associated factors.

results and findings from a seminal twin study that estimated educational attainment heritability by birth cohort¹⁶).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 June 2015; accepted 16 March 2016.

Published online 11 May 2016.

1. Rietveld, C. A. *et al.* GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340**, 1467–1471 (2013).
2. Rietveld, C. A. *et al.* Replicability and robustness of genome-wide-association studies for behavioral traits. *Psychol. Sci.* **25**, 1975–1986 (2014).
3. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).
4. Bulik-Sullivan, B. K. *et al.* LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
5. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
6. Fowler, T., Zammit, S., Owen, M. J. & Rasmussen, F. A population-based study of shared genetic variation between premorbid IQ and psychosis among male twin pairs and sibling pairs from Sweden. *Arch. Gen. Psychiatry* **69**, 460–466 (2012).
7. Tambs, K., Sundet, J. M., Magnus, P. & Berg, K. Genetic and environmental contributions to the covariance between occupational status, educational attainment, and IQ: a study of twins. *Behav. Genet.* **19**, 209–222 (1989).
8. Thompson, L. A., Dettmerman, D. K. & Plomin, R. Associations between cognitive abilities and scholastic achievement: Genetic overlap but environmental differences. *Psychol. Sci.* **2**, 158–165 (1991).
9. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
10. Ardlie, K. G. *et al.*; GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
11. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
12. Allen Institute for Brain Science. BrainSpan atlas of the developing human brain <http://www.brainspan.org> (2015).
13. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
14. Krapohl, E. *et al.* The high heritability of educational achievement reflects many genetically influenced traits, not just intelligence. *Proc. Natl Acad. Sci. USA* **111**, 15273–15278 (2014).

15. Branigan, A. R., McCallum, K. J. & Freese, J. Variation in the heritability of educational attainment: An international meta-analysis. *Social Forces* **92**, 109–140 (2013).
16. Heath, A. C. *et al.* Education policy and the heritability of educational attainment. *Nature* **314**, 734–736 (1985).
17. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genetics* **47**, 1228–1235 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements This research was carried out under the auspices of the Social Science Genetic Association Consortium (SSGAC). This research has also been conducted using the UK Biobank Resource. This study was supported by funding from the Ragnar Söderberg Foundation (E9/11), the Swedish Research Council (421-2013-1061), The Jan Wallander and Tom Hedelius Foundation, an ERC Consolidator Grant (647648 EdGe), the Pershing Square Fund of the Foundations of Human Behavior, and the NIA/NIH through grants P01-AG005842, P01-AG005842-20S2, P30-AG012810, and T32-AG000186-23 to NBER, and R01-AG042568 to USC. We thank S. Cunningham, N. Galla and J. Rashtian for research assistance. A full list of acknowledgments is provided in the Supplementary Information.

Author Contributions Study design and management: D.J.B., D.Ce., T.E., M.J., P.D.K. and P.M.V. Quality control and meta-analysis: A.O., G.B.C., T.E., M.A.F., C.A.R. and T.H.P. Stratification: P.T., J.P.B., C.A.R. and J.Y. Genetic overlap: J.P.B., M.A.F., P.T. Biological annotation: J.J.L., T.E., T.H.P., J.K.P., J.H.B., J.P.B., L.F., V.E., G.A.M., M.A.F., S.F.W.M., P.Ti., R.A.P., R.d.V. and H.J.W. Prediction and mediation: J.P.B., M.A.F. and J.Y. G×E: D.Co., S.F.L., K.O.L., S.O. and K.T. Replication in UKB: M.A.F. and C.A.R. SSGAC advisory board: D.Co., T.E., A.H., R.F.K., D.I.L., S.E.M., M.N.M., G.D.S. and P.M.V. All authors contributed to and critically reviewed the manuscript. Authors not listed above contributed to the recruitment, genotyping, or data processing for the contributing components of the meta-analysis. For a full list of author contributions, see Supplementary Information section 8.

Author Information Results can be downloaded from the SSGAC website (<http://ssgac.org/Data.php>). Data for our analyses come from many studies and organizations, some of which are subject to a MTA, and are listed in the Supplementary Information. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.J.B. (daniel.benjamin@gmail.com), D.Ce. (dac12@nyu.edu), P.D.K. (p.d.koellinger@vu.nl) or P.M.V. (peter.visscher@uq.edu.au).

Aysu Okbay^{1,2,3*}, Jonathan P. Beauchamp^{4*}, Mark Alan Fontana^{5*}, James J. Lee^{6*}, Tine H. Pers^{7,8,9,10*}, Cornelius A. Rietveld^{1,2,3*}, Patrick Turley^{4*}, Guo-Bo Chen¹¹, Valur Emilsson^{12,13}, S. Fleur W. Meddens^{3,14,15}, Sven Oskarsson¹⁶, Joseph K. Pickrel¹⁷, Kevin Thom¹⁸, Pascal Timshe^{8,19}, Ronald de Vlaming^{1,2,3}, Abdel Abdellaoui²⁰, Tarunveer S. Ahluwalia^{9,21,22}, Jonas Bacelis²³, Clemens Baumbach^{24,25}, Gyda Bjornsdottir²⁶, Johannes H. Brandma²⁷, Maria Pina Concas²⁸, Jaime Derringer²⁹, Nicholas A. Furlotte³⁰, Tessel E. Galesloot³¹, Giorgia Grotto³², Richa Gupta³³, Leanne M. Hall^{34,35}, Sarah E. Harris^{36,37}, Edith Hofer^{38,39}, Momoko Horikoshi^{40,41}, Jennifer E. Huffman⁴², Kadri Kaasik⁴³, Ioanna P. Kalafati⁴⁴, Robert Karlsson⁴⁵, Augustine Kong²⁶, Jari Lahti^{43,46}, Sven J. van der Lee², Christiaan de Leeuw^{14,47}, Penelope A. Lind⁴⁸, Karl-Oskar Lindgren⁶⁰, Tian Liu⁴⁹, Massimo Mangino^{50,51}, Jonathan Marten⁴², Evelin Mihailov⁵², Michael B. Miller⁶, Peter J. van der Most⁵³, Christopher Oldmeadow^{54,55}, Antony Payton^{56,57}, Natalia Pervjakova^{52,58}, Wouter J. Peyrot⁵⁹, Yong Qian⁶⁰, Olli Raitakari⁶¹, Rico Rueedi^{52,63}, Erika Salvi⁶⁴, Borge Schmidt⁶⁵, Katharina E. Schraut⁶⁶, Jianxin Shi⁶⁷, Albert V. Smith^{68,69}, Raymond A. Poot²⁷, Beate St Pourcain^{70,71}, Alexander Teumer⁷², Gudmar Thorleifsson²⁶, Nick Verweij⁷³, Dragana Vuckovic⁷², Juergen Wellmann⁷⁴, Harm-Jan Westra^{8,75,76}, Jingyun Yang^{77,78}, Wei Zhao⁷⁹, Zhihong Zhu¹¹, Behrooz Z. Alizadeh^{83,80}, Najaf Amin², Andrew Bakshi¹¹, Sebastian E. Baumeister^{72,81}, Ginevra Biino⁸², Klaus Bonnelykke²¹, Patricia A. Boyle^{77,83}, Harry Campbell⁶⁶, Francesco P. Cappucco⁸⁴, Gail Davies^{36,85}, Jan-Emmanuel De Neve⁸⁶, Panos Deloukas^{87,88}, Ilya Demuth^{89,90}, Jun Ding⁶⁰, Peter Eibich^{91,92}, Lewin Eisele⁶⁵, Niina Eklund⁵⁸, David M. Evans^{70,93}, Jessica D. Faul⁹⁴, Mary F. Feitosa⁹⁵, Andreas J. Forstner^{96,97}, Ilmaria Gandin³², Bjarni Gunnarsson²⁶, Bjarni V. Halldorsson^{26,98}, Tamara B. Harris⁹⁹, Andrew C. Heath¹⁰⁰, Lynne J. Hocking¹⁰¹, Elizabeth G. Holliday^{54,55}, Georg Homuth¹⁰², Michael A. Horan¹⁰³, Jouke-Jan Hottenga⁴⁰, Philip L. de Jager^{8,104,105}, Peter K. Joshi⁶⁶, Astanand Jugessur¹⁰⁶, Marika A. Kaakinen¹⁰⁷, Mika Kahonen^{108,109}, Stavroula Kanoni⁸⁷, Liisa Keltigangas-Järvinen⁴³, Lambertus A. L. M. Kiemeny³¹, Ivana Kolcic¹¹⁰, Seppo Koskinen⁵⁸, Aldi T. Kraja⁹⁵, Martin Kroh⁹¹, Zoltan Kutalik^{62,63,111}, Antti Latvala³³, Lenore J. Launer¹¹², Maël P. Lebreton^{115,113}, Douglas F. Levinson¹¹⁴, Paul Lichtenstein⁴⁵, Peter Lichtner¹¹⁵, David C. M. Liewald^{36,85}, LifeLines Cohort Study[†], Anu Loukola³³, Pamela A. Madden¹⁰⁰, Reedik Mägi⁵², Tomi Mäki-Opas⁵⁸, Riccardo E. Marioni^{11,36,116}, Pedro Marques-Vidal¹¹⁷, Gerardus A. Meddens¹¹⁸, George McMahon⁷⁰, Christa Meisinger²⁵, Thomas Meitinger¹¹⁵, Yusupliri Milaneschi⁵⁹, Lili Milani⁵², Grant W. Montgomery¹¹⁹, Ronny Myhre¹⁰⁶, Christopher P. Nelson^{24,35}, Dale R. Nyholt^{119,120}, William E. R. Ollier⁵⁶, Aarno Palotie^{8,121,122,123,124,125}, Lavinia Paternoster⁷⁰, Nancy L. Pedersen⁴⁵, Katja E. Petrovic³⁸, David J. Porteous³⁷, Katri Räikkönen^{43,46}, Susan M. Ring⁷⁰, Antonietta Robino¹²⁶, Olga Rostapshova^{4,127}, Igor Rudan⁶⁶, Aldo Rustichini¹²⁸, Veikko Salomaa⁵⁸, Alan R. Sanders^{129,130}, Antti-Pekka Sarin^{124,131}, Helena Schmidt^{38,132}, Rodney J. Scott^{55,133}, Blair H. Smith¹³⁴, Jennifer A. Smith⁷⁹, Jan A. Staessen^{135,136}, Elisabeth Steinhagen-Thiessen⁸⁹, Konstantin Strauch^{137,138}, Antonio Terracciano¹³⁹, Martin D. Tobin¹⁴⁰, Sheila Ulivi¹²⁶, Simona Vaccargiu²⁸, Lydia Quayle⁵⁰, Frank J. A. van Rooij^{52,141}, Cristina Venturini^{50,51}, Anna A. E. Vinkhuyzen¹¹, Uwe Völker¹⁰², Henry Völzke⁷², Judith M. Vonk⁵³, Diego Vozzi¹²⁷, Johannes Waage^{21,22}, Erin B. Ware^{79,142}, Gonneke Willemssen²⁰, John R. Attia^{54,55}, David A. Bennett^{77,78}, Klaus Berger⁷³, Lars Bertram^{143,144}, Hans Bisgaard²¹, Dorret I. Boomsma²⁰, Ingrid B. Borecki⁹⁵, Ute Bültmann¹⁴⁵, Christopher F. Chabris¹⁴⁶, Francesco Cucca¹⁴⁷, Daniele Cusi^{64,148}, Ian J. Deary^{36,85}, George V. Dedoussis⁴⁴, Cornelia M. van Duijn², Johan G. Eriksson^{46,149}, Barbara Franke¹⁵⁰, Lude Franke¹⁵¹, Paolo Gasparin^{32,126,152}, Pablo V. Gejman^{129,130}, Christian Gieger²⁴, Hans-Jörgen Grabe^{153,154}, Jacob Gratten¹¹, Patrick J. F. Groenen¹⁵⁵, Vilhelmur Gudnason^{12,69}, Pim van der Harst^{73,151,156}, Caroline Hayward^{42,157}, David A. Hinds³⁰, Wolfgang Hoffmann⁷², Elina Hyppönen^{158,159,160}, William G. Iacono⁶, Bo Jacobsson^{23,106}, Marjo-Riitta Järvelin^{161,162,163,164}, Karl-Heinz Jöckel⁶⁵, Jaakko Kaprio^{33,58,124}, Sharon L. R. Kardia⁷⁹, Terho Lehtimäki^{165,166}, Steven F. Lehrer^{167,168}, Patrik K. E. Magnusson⁴⁵, Nicholas G. Martin¹⁶⁹, Matt McGue⁶, Andres Metspalu^{52,170}, Neil Pendleton^{71,172}, Brenda W. J. H. Penninx⁵⁹, Markus Perola^{52,58}, Nicola Pirastu³², Mario Pirastu²⁸, Ozren Polasek^{66,173}, Danielle Posthuma^{14,174}, Christine Power¹⁶⁰, Michael A. Province⁹⁵, Nilesh J. Samani^{34,35}, David Schlessinger⁶⁰, Reinhold Schmidt³⁸, Thorkild I. A. Sørensen^{9,70,175}, Tim D. Spector⁵⁰, Kari Stefansson^{26,69}, Unnur Thorsteinsdottir^{26,69}, A. Roy Thurik^{1,3,176,177}, Nicholas J. Timpson⁷⁰, Henning Tiemeier^{2,178,179}, Joyce Y. Tung³⁰, André G. Uitterlinden^{2,180}, Veronique Vitart⁴², Peter Vollenweider¹¹⁷, David R. Weir⁹⁴, James F. Wilson^{42,66}, Alan F. Wright⁴², Dalton C. Conley^{181,182}, Robert F. Krueger⁶, George Davey Smith⁷⁰, Albert Hofman², David I. Laibson⁴, Sarah E. Medland⁴⁸, Michelle N. Meyer¹⁸³, Jian Yang^{11,93}, Magnus Johannesson¹⁸⁴, Peter M. Visscher^{11,93}, Tõnu Esko^{7,8,52,185}, Philipp D. Koellinger^{3,14,15}, David Cesarini^{18,186} & Daniel J. Benjamin⁵

¹Department of Applied Economics, Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, 3062 PA, The Netherlands. ²Department of Epidemiology, Erasmus Medical Center, Rotterdam, 3015 GE, The Netherlands. ³Erasmus University Rotterdam Institute for Behavior and Biology, Rotterdam, 3062 PA, The Netherlands. ⁴Department of Economics, Harvard University, Cambridge, Massachusetts 02138, USA. ⁵Center for Economic and Social Research, University of Southern California, Los Angeles, California 90089-3332, USA. ⁶Department of

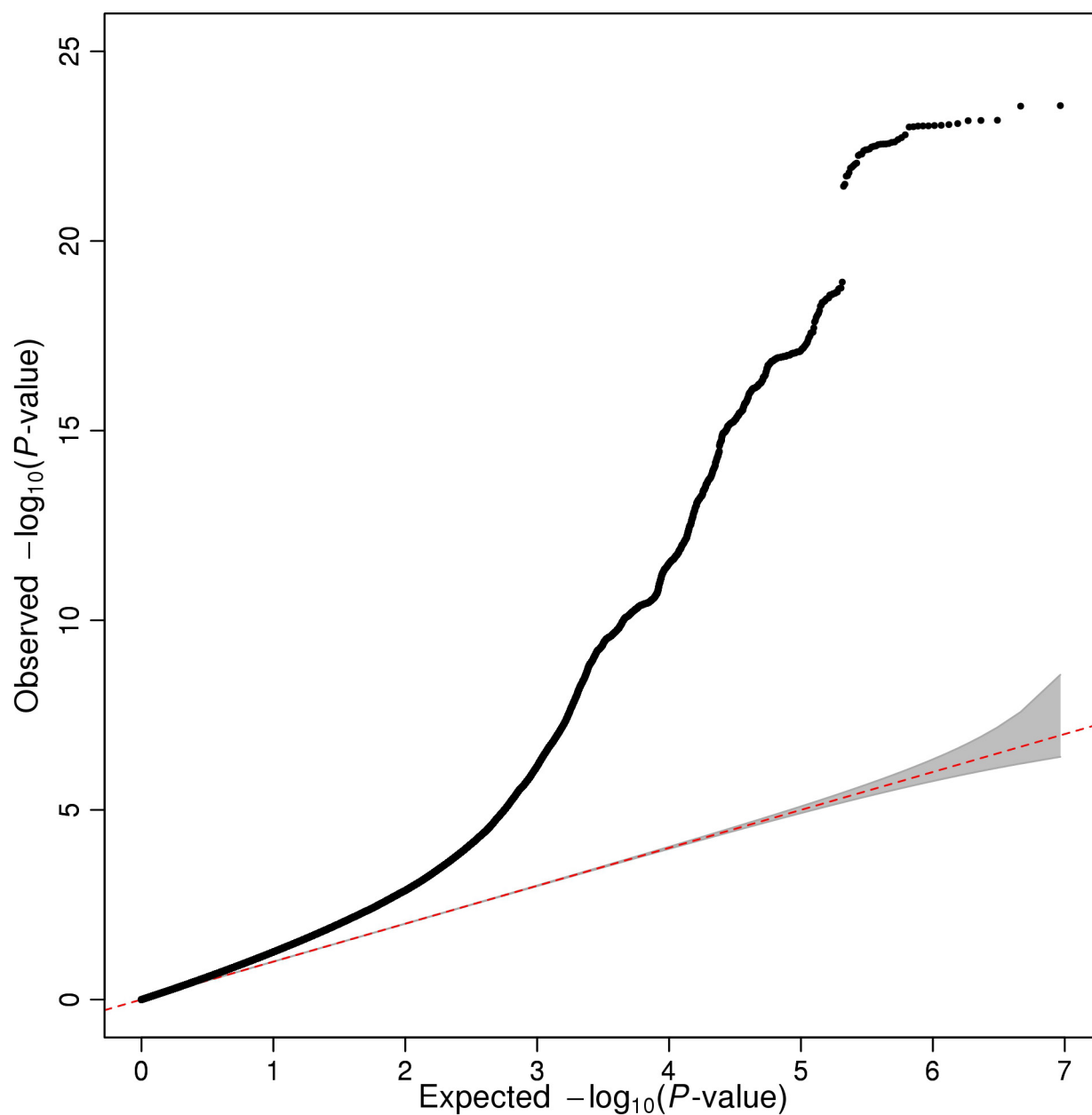
Psychology, University of Minnesota Twin Cities, Minneapolis, Minnesota 55455, USA. ⁷Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, Massachusetts 2116, USA. ⁸Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. ⁹The Novo Nordisk Foundation Center for Basic Metabolic Research, Section of Metabolic Genetics, University of Copenhagen, Faculty of Health and Medical Sciences, Copenhagen 2100, Denmark. ¹⁰Statens Serum Institut, Department of Epidemiology Research, Copenhagen 2300, Denmark. ¹¹Queensland Brain Institute, The University of Queensland, Brisbane, QLD 4072, Australia. ¹²Icelandic Heart Association, Kopavogur 201, Iceland. ¹³Faculty of Pharmaceutical Sciences, University of Iceland, Reykjavik 107, Iceland. ¹⁴Department of Complex Trait Genetics, VU University, Center for Neurogenetics and Cognitive Research, Amsterdam, 1081 HV, The Netherlands. ¹⁵Amsterdam Business School, University of Amsterdam, Amsterdam, 1018 TV, The Netherlands. ¹⁶Department of Government, Uppsala University, Uppsala 751 20, Sweden. ¹⁷New York Genome Center, New York, New York 10013, USA. ¹⁸Department of Economics, New York University, New York, New York 10012, USA. ¹⁹Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark Lyngby 2800, Denmark. ²⁰Department of Biological Psychology, VU University Amsterdam, Amsterdam, 1081 BT, The Netherlands. ²¹COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Herlev and Gentofte Hospital, University of Copenhagen, Copenhagen 2820, Denmark. ²²Steno Diabetes Center, Gentofte 2820, Denmark. ²³Department of Obstetrics and Gynecology, Institute of Clinical Sciences, Sahlgrenska Academy, Gothenburg 416 85, Sweden. ²⁴Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg 85764, Germany. ²⁵Institute of Epidemiology II, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg 85764, Germany. ²⁶deCODE Genetics/Amgen Inc., Reykjavik 101, Iceland. ²⁷Department of Cell Biology, Erasmus Medical Center Rotterdam, 3015 CN, The Netherlands. ²⁸Istituto di Ricerca Genetica e Biomedica U.O.S. di Sassari, National Research Council of Italy, Sassari 07100, Italy. ²⁹Psychology, University of Illinois, Champaign, Illinois 61820, USA. ³⁰23andMe, Inc., Mountain View, California 94041, USA. ³¹Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, 6500 HB, The Netherlands. ³²Department of Medical, Surgical and Health Sciences, University of Trieste, Trieste 34100, Italy. ³³Department of Public Health, University of Helsinki, 00014 Helsinki, Finland. ³⁴Department of Cardiovascular Sciences, University of Leicester, Leicester LE3 9QP, UK. ³⁵NIHR Leicester Cardiovascular Biomedical Research Unit, Glenfield Hospital, Leicester LE3 9QP, UK. ³⁶Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh EH8 9JZ, UK. ³⁷Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK. ³⁸Department of Neurology, General Hospital and Medical University Graz, Graz 8036, Austria. ³⁹Institute for Medical Informatics, Statistics and Documentation, General Hospital and Medical University Graz, Graz 8036, Austria. ⁴⁰Oxford Centre for Diabetes, Endocrinology & Metabolism, University of Oxford, Oxford OX3 7LE, UK. ⁴¹Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. ⁴²MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK. ⁴³Institute of Behavioural Sciences, University of Helsinki, 00014 Helsinki, Finland. ⁴⁴Nutrition and Dietetics, Health Science and Education, Harokopio University, Athens 17671, Greece. ⁴⁵Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm 171 77, Sweden. ⁴⁶Folkhälsan Research Centre, 00014 Helsingfors, Finland. ⁴⁷Institute for Computing and Information Sciences, Radboud University Nijmegen, Nijmegen, 6525 EC, The Netherlands. ⁴⁸Quantitative Genetics, QIMR Berghofer Medical Research Institute, Brisbane, QLD 4029, Australia. ⁴⁹Lifespan Psychology, Max Planck Institute for Human Development, Berlin 14195, Germany. ⁵⁰Department of Twin Research and Genetic Epidemiology, King's College London, London SE1 7EH, UK. ⁵¹NIHR Biomedical Research Centre, Guy's and St. Thomas' Foundation Trust, London SE1 7EH, UK. ⁵²Estonian Genome Center, University of Tartu, Tartu 51010, Estonia. ⁵³Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, 9700 RB, The Netherlands. ⁵⁴Public Health Stream, Hunter Medical Research Institute, New Lambton, NSW 2305, Australia. ⁵⁵Faculty of Health and Medicine, University of Newcastle, Newcastle, NSW 2300, Australia. ⁵⁶Centre for Integrated Genomic Medical Research, Institute of Population Health, The University of Manchester, Manchester M13 9PT, UK. ⁵⁷Human Communication and Deafness, School of Psychological Sciences, The University of Manchester, Manchester M13 9PL, UK. ⁵⁸Department of Health, THL-National Institute for Health and Welfare, 00271 Helsinki, Finland. ⁵⁹Psychiatry, VU University Medical Center & GGZ inGeest, Amsterdam, 1081 HL, The Netherlands. ⁶⁰Laboratory of Genetics, National Institute on Aging, Baltimore, Maryland 21224, USA. ⁶¹Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, 20521 Turku, Finland. ⁶²Department of Medical Genetics, University of Lausanne, Lausanne 1005, Switzerland. ⁶³Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland. ⁶⁴Department of Health Sciences, University of Milan, Milano 20142, Italy. ⁶⁵Institute for Medical Informatics, Biometry and Epidemiology, University Hospital of Essen, Essen 45147, Germany. ⁶⁶Centre for Global Health Research, The Usher Institute for Population Health Sciences and Informatics, University of Edinburgh, Edinburgh EH8 9AG, UK. ⁶⁷Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland 20892-9780, USA. ⁶⁸Icelandic Heart Association, Kopavogur 201, Iceland. ⁶⁹Faculty of Medicine, University of Iceland, Reykjavik 101, Iceland. ⁷⁰MRC Integrative Epidemiology Unit, University of Bristol, Bristol BS8 2BN, UK. ⁷¹School of Oral and Dental Sciences, University of Bristol, Bristol BS1 2LY, UK. ⁷²Institute for Community Medicine, University Medicine Greifswald, Greifswald 17475, Germany. ⁷³Department of Cardiology, University

- Medical Center Groningen, University of Groningen, Groningen, 9700 RB, The Netherlands. ⁷⁴Institute of Epidemiology and Social Medicine, University of Münster, Münster 48149, Germany. ⁷⁵Divisions of Genetics and Rheumatology, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. ⁷⁶Partners Center for Personalized Genetic Medicine, Boston, Massachusetts 02115, USA. ⁷⁷Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, Illinois 60612, USA. ⁷⁸Department of Neurological Sciences, Rush University Medical Center, Chicago, Illinois 60612, USA. ⁷⁹Department of Epidemiology, University of Michigan, Ann Arbor, Michigan 48109, USA. ⁸⁰Department of Gastroenterology and Hepatology, University of Groningen, University Medical Center Groningen, Groningen, 9713 GZ, The Netherlands. ⁸¹Institute of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg D-93053, Germany. ⁸²Institute of Molecular Genetics, National Research Council of Italy, Pavia 27100, Italy. ⁸³Department of Behavioral Sciences, Rush University Medical Center, Chicago, Illinois 60612, USA. ⁸⁴Warwick Medical School, University of Warwick, Coventry CV4 7AL, UK. ⁸⁵Department of Psychology, University of Edinburgh, Edinburgh EH8 9JZ, UK. ⁸⁶Said Business School, University of Oxford, Oxford OX1 1HP, UK. ⁸⁷William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London EC1M 6BQ, UK. ⁸⁸Princess Al-Jawhara Al-Brahim Centre of Excellence in Research of Hereditary Disorders (PACER-HD), King Abdulaziz University, Jeddah 21589, Saudi Arabia. ⁸⁹The Berlin Aging Study II; Research Group on Geriatrics, Charité – Universitätsmedizin Berlin, Germany, Berlin 13347, Germany. ⁹⁰Institute of Medical and Human Genetics, Charité-Universitätsmedizin, Berlin, Berlin 13353, Germany. ⁹¹German Socio-Economic Panel Study, DIW Berlin, Berlin 10117, Germany. ⁹²Health Economics Research Centre, Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, UK. ⁹³The University of Queensland Diamantina Institute, The Translational Research Institute, Brisbane, QLD 4102, Australia. ⁹⁴Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan 48109, USA. ⁹⁵Department of Genetics, Division of Statistical Genomics, Washington University School of Medicine, St. Louis, Missouri 63018, USA. ⁹⁶Institute of Human Genetics, University of Bonn, Bonn 53127, Germany. ⁹⁷Department of Genomics, Life and Brain Center, University of Bonn, Bonn 53127, Germany. ⁹⁸Institute of Biomedical and Neural Engineering, School of Science and Engineering, Reykjavik University, Reykjavik 101, Iceland. ⁹⁹Laboratory of Epidemiology, Demography, National Institute on Aging, National Institutes of Health, Bethesda, Maryland 20892-9205, USA. ¹⁰⁰Department of Psychiatry, Washington University School of Medicine, St. Louis, Missouri 63110, USA. ¹⁰¹Division of Applied Health Sciences, University of Aberdeen, Aberdeen AB25 2ZD, UK. ¹⁰²Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald 17475, Germany. ¹⁰³Manchester Medical School, The University of Manchester, Manchester M13 9PT, UK. ¹⁰⁴Program in Translational NeuroPsychiatric Genomics, Departments of Neurology & Psychiatry, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. ¹⁰⁵Harvard Medical School, Boston, Massachusetts 02115, USA. ¹⁰⁶Department of Genes and Environment, Norwegian Institute of Public Health, N-0403 Oslo, Norway. ¹⁰⁷Department of Genomics of Common Disease, Imperial College London, London, W12 0NN, UK. ¹⁰⁸Department of Clinical Physiology, Tampere University Hospital, 33521 Tampere, Finland. ¹⁰⁹Department of Clinical Physiology, University of Tampere, School of Medicine, 33014 Tampere, Finland. ¹¹⁰Public Health, Medical School, University of Split, 21000 Split, Croatia. ¹¹¹Institute of Social and Preventive Medicine, Lausanne University Hospital (CHUV), Lausanne 1010, Switzerland. ¹¹²Neuroepidemiology Section, National Institute on Aging, National Institutes of Health, Bethesda, Maryland 20892-9205, USA. ¹¹³Amsterdam Brain and Cognition Center, University of Amsterdam, Amsterdam, 1018 XA, The Netherlands. ¹¹⁴Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, California 94305-5797, USA. ¹¹⁵Institute of Human Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg 85764, Germany. ¹¹⁶Medical Genetics Section, Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, EH4 2XU, UK. ¹¹⁷Department of Internal Medicine, Internal Medicine, Lausanne University Hospital (CHUV), Lausanne 1011, Switzerland. ¹¹⁸Tema BV, Hoofddorp, 2131 HE, The Netherlands. ¹¹⁹Molecular Epidemiology, QIMR Berghofer Medical Research Institute, Brisbane, QLD 4029, Australia. ¹²⁰Institute of Health and Biomedical Innovation, Queensland Institute of Technology, Brisbane, QLD 4059, Australia. ¹²¹Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. ¹²²The Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. ¹²³Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. ¹²⁴Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki 00014, Finland. ¹²⁵Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. ¹²⁶Medical Genetics, Institute for Maternal and Child Health IRCCS "Burlo Garofolo", Trieste 34100, Italy. ¹²⁷Social Impact, Arlington, Virginia 22201, USA. ¹²⁸Department of Economics, University of Minnesota Twin Cities, Minneapolis, Minnesota 55455, USA. ¹²⁹Department of Psychiatry and Behavioral Sciences, NorthShore University HealthSystem, Evanston, Illinois 60201-3137, USA. ¹³⁰Department of Psychiatry and Behavioral Neuroscience, University of Chicago, Chicago, Illinois 60637, USA. ¹³¹Public Health Genomics Unit, National Institute for Health and Welfare, 00300 Helsinki, Finland. ¹³²Research Unit for Genetic Epidemiology, Institute of Molecular Biology and Biochemistry, Center of Molecular Medicine, General Hospital and Medical University, Graz, Graz 8010, Austria. ¹³³Information Based Medicine Stream, Hunter Medical Research Institute, New Lambton, NSW 2305, Australia. ¹³⁴Medical Research Institute, University of Dundee, Dundee DD1 9SY, UK. ¹³⁵Research Unit Hypertension and Cardiovascular Epidemiology, Department of Cardiovascular Science, University of Leuven, Leuven 3000, Belgium. ¹³⁶R&D VitaK Group, Maastricht University, Maastricht, 6229 EV, The Netherlands. ¹³⁷Institute of Genetic Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg 85764, Germany. ¹³⁸Institute of Medical Informatics, Biometry and Epidemiology, Chair of Genetic Epidemiology, Ludwig Maximilians-Universität, Munich 81377, Germany. ¹³⁹Department of Geriatrics, Florida State University College of Medicine, Tallahassee, Florida 32306, USA. ¹⁴⁰Department of Health Sciences and Genetics, University of Leicester, Leicester LE1 7RH, UK. ¹⁴¹Department of Internal Medicine, Erasmus Medical Center, Rotterdam, 3015 GE, The Netherlands. ¹⁴²Research Center for Group Dynamics, Institute for Social Research, University of Michigan, Ann Arbor, Michigan 48104, USA. ¹⁴³Platform for Genome Analytics, Institutes of Neurogenetics & Integrative and Experimental Genomics, University of Lübeck, Lübeck 23562, Germany. ¹⁴⁴Neuroepidemiology and Ageing Research Unit, School of Public Health, Faculty of Medicine, Imperial College of Science, Technology and Medicine, London SW7 2AZ, UK. ¹⁴⁵Department of Health Sciences, Community & Occupational Medicine, University of Groningen, University Medical Center Groningen, Groningen, 9713 AV, The Netherlands. ¹⁴⁶Department of Psychology, Union College, Schenectady, New York 12308, USA. ¹⁴⁷Istituto di Ricerca Genetica e Biomedica (IRGB), Consiglio Nazionale delle Ricerche, c/o Cittadella Universitaria di Monserrato, Monserrato, Cagliari 9042, Italy. ¹⁴⁸Institute of Biomedical Technologies, Italian National Research Council, Segrate (Milano) 20090, Italy. ¹⁴⁹Department of General Practice and Primary Health Care, University of Helsinki, 00014 Helsinki, Finland. ¹⁵⁰Departments of Human Genetics and Psychiatry, Donders Centre for Neuroscience, Nijmegen, 6500 HB, The Netherlands. ¹⁵¹Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, 9700 RB, The Netherlands. ¹⁵²Sidra, Experimental Genetics Division, Sidra, Doha 26999, Qatar. ¹⁵³Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Greifswald 17475, Germany. ¹⁵⁴Department of Psychiatry and Psychotherapy, HELIOS-Hospital Stralsund, Stralsund 18437, Germany. ¹⁵⁵Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, 3062 PA, The Netherlands. ¹⁵⁶Durrer Center for Cardiogenetic Research, ICIN-Netherlands Heart Institute, Utrecht, 1105 AZ, The Netherlands. ¹⁵⁷Generation Scotland, Centre for Genomics and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK. ¹⁵⁸Centre for Population Health Research, School of Health Sciences and Sansom Institute, University of South Australia, Adelaide, SA 5000, Australia. ¹⁵⁹South Australian Health and Medical Research Institute, Adelaide, SA 5000, Australia. ¹⁶⁰Population, Policy and Practice, UCL Institute of Child Health, London WC1N 1EH, UK. ¹⁶¹Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment & Health, School of Public Health, Imperial College London, London W2 1PG, UK. ¹⁶²Center for Life Course Epidemiology, Faculty of Medicine, University of Oulu, 90014 Oulu, Finland. ¹⁶³Unit of Primary Care, Oulu University Hospital, 90029 Oulu, Finland. ¹⁶⁴Biocenter Oulu, University of Oulu, 90014 Oulu, Finland. ¹⁶⁵Fimlab Laboratories, 33520 Tampere, Finland. ¹⁶⁶Department of Clinical Chemistry, University of Tampere, School of Medicine, 33014 Tampere, Finland. ¹⁶⁷Economics, NYU Shanghai, 200122 Pudong, China. ¹⁶⁸Policy Studies, Queen's University, Kingston, Ontario K7L 3N6, Canada. ¹⁶⁹Genetic Epidemiology, QIMR Berghofer Medical Research Institute, Brisbane, QLD 4029, Australia. ¹⁷⁰Institute of Molecular and Cell Biology, University of Tartu, Tartu 51010, Estonia. ¹⁷¹Centre for Clinical and Cognitive Neuroscience, Institute Brain Behaviour and Mental Health, Salford Royal Hospital, Manchester M6 8HD, UK. ¹⁷²Manchester Institute for Collaborative Research in Ageing, University of Manchester, Manchester M13 9PL, UK. ¹⁷³Faculty of Medicine, University of Split, Split 21000, Croatia. ¹⁷⁴Department of Clinical Genetics, VU Medical Centre, Amsterdam, 1081 HV, The Netherlands. ¹⁷⁵Institute of Preventive Medicine, Bispebjerg and Frederiksberg Hospitals, The Capital Region, Frederiksberg 2000, Denmark. ¹⁷⁶Montpellier Business School, Montpellier 34080, France. ¹⁷⁷Panteia, Zoetermeer, 2715 CA, The Netherlands. ¹⁷⁸Department of Psychiatry, Erasmus Medical Center, Rotterdam, 3015 GE, The Netherlands. ¹⁷⁹Department of Child and Adolescent Psychiatry, Erasmus Medical Center, Rotterdam, 3015 GE, The Netherlands. ¹⁸⁰Department of Internal Medicine, Erasmus Medical Center, Rotterdam, 3015 GE, The Netherlands. ¹⁸¹Department of Sociology, New York University, New York, New York 10012, USA. ¹⁸²School of Medicine, New York University, New York, New York 10016, USA. ¹⁸³Bioethics Program, Union Graduate College – Icahn School of Medicine at Mount Sinai, Schenectady, New York 12308, USA. ¹⁸⁴Department of Economics, Stockholm School of Economics, Stockholm 113 83, Sweden. ¹⁸⁵Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. ¹⁸⁶Research Institute for Industrial Economics, Stockholm 10215, Sweden.

*These authors contributed equally to this work.

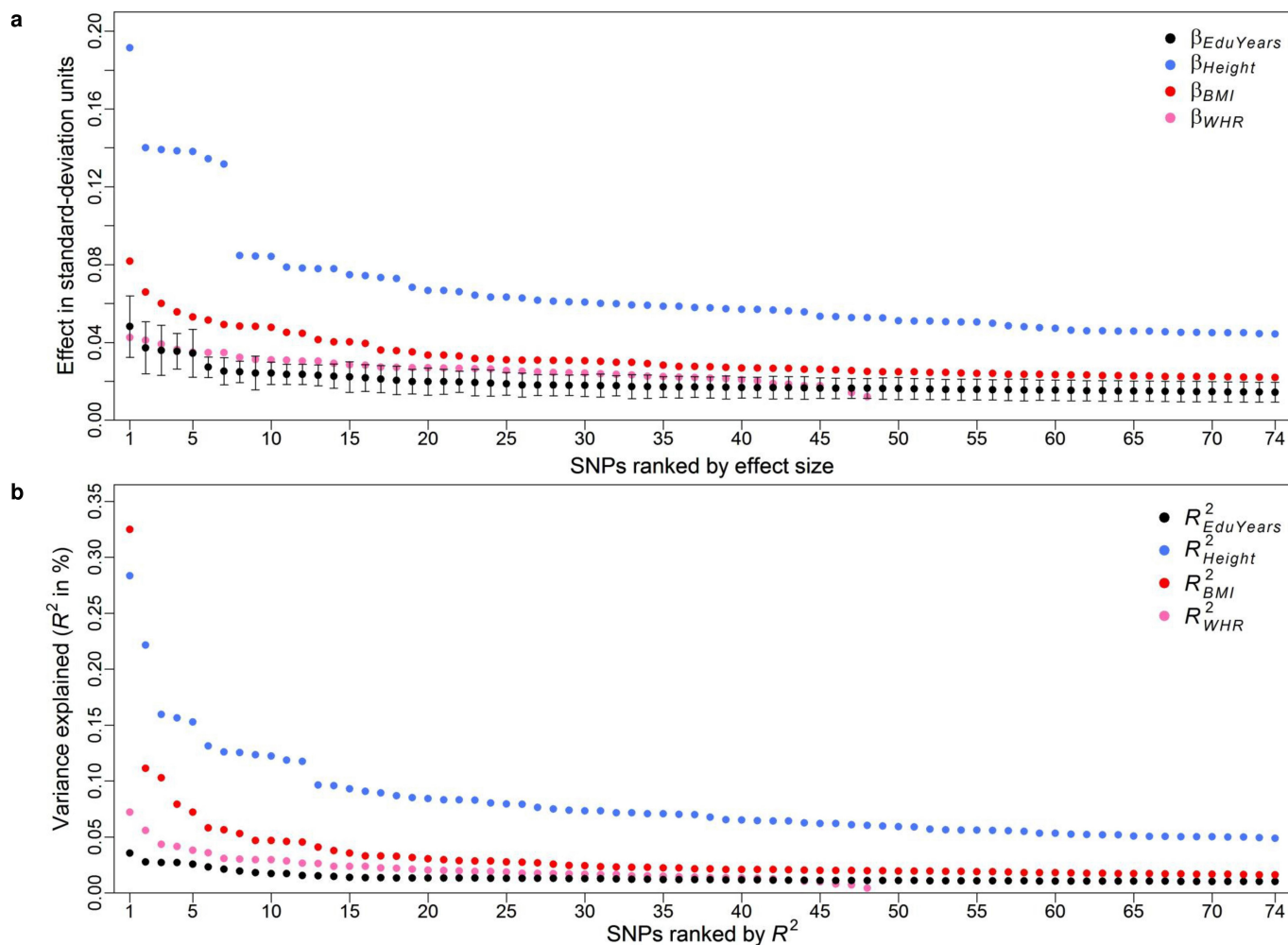
§These authors jointly supervised this work.

†A list of participants and affiliations appears in the Supplementary Information.



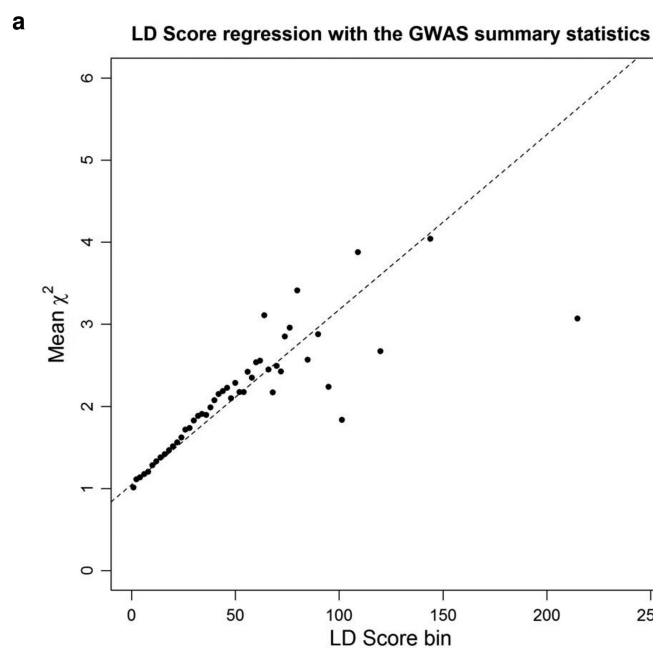
Extended Data Figure 1 | Q-Q plot of the genome-wide association meta-analysis of 64 EduYears results files ($n = 293,723$). Observed and expected P values are on a $-\log_{10}$ scale (two-tailed). The grey region depicts the 95% confidence interval under the null hypothesis of

a uniform P value distribution. The observed λ_{GC} is 1.28. (As reported in Supplementary Information section 1.5.4, the unweighted mean λ_{GC} is 1.02, the unweighted median is 1.01, and the range across cohorts is 0.95–1.15.)

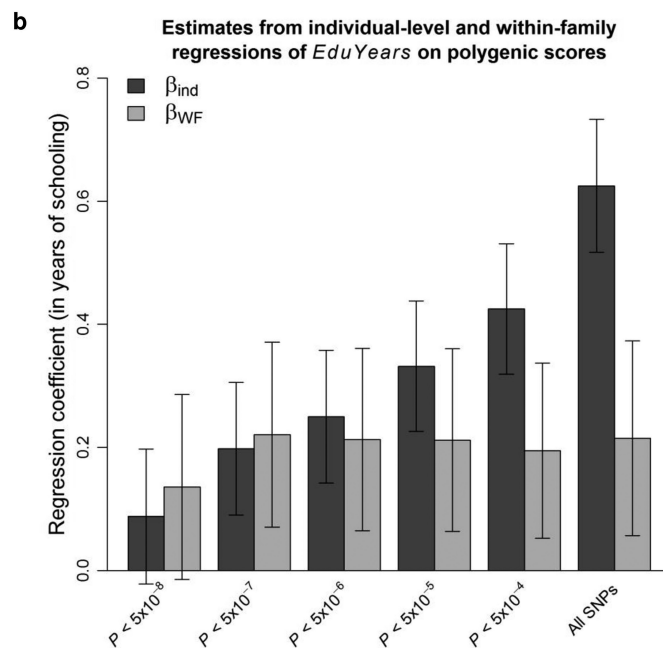


Extended Data Figure 2 | The distribution of effect sizes of the 74 lead SNPs. **a**, SNPs ordered by absolute value of the standardized effect of one more copy of the education-increasing allele, with 95% confidence intervals. **b**, SNPs ordered by R^2 . Effects on EduYears are benchmarked against the top 74 genome-wide significant hits identified in the largest GWAS conducted to date of height and body mass index (BMI), and the

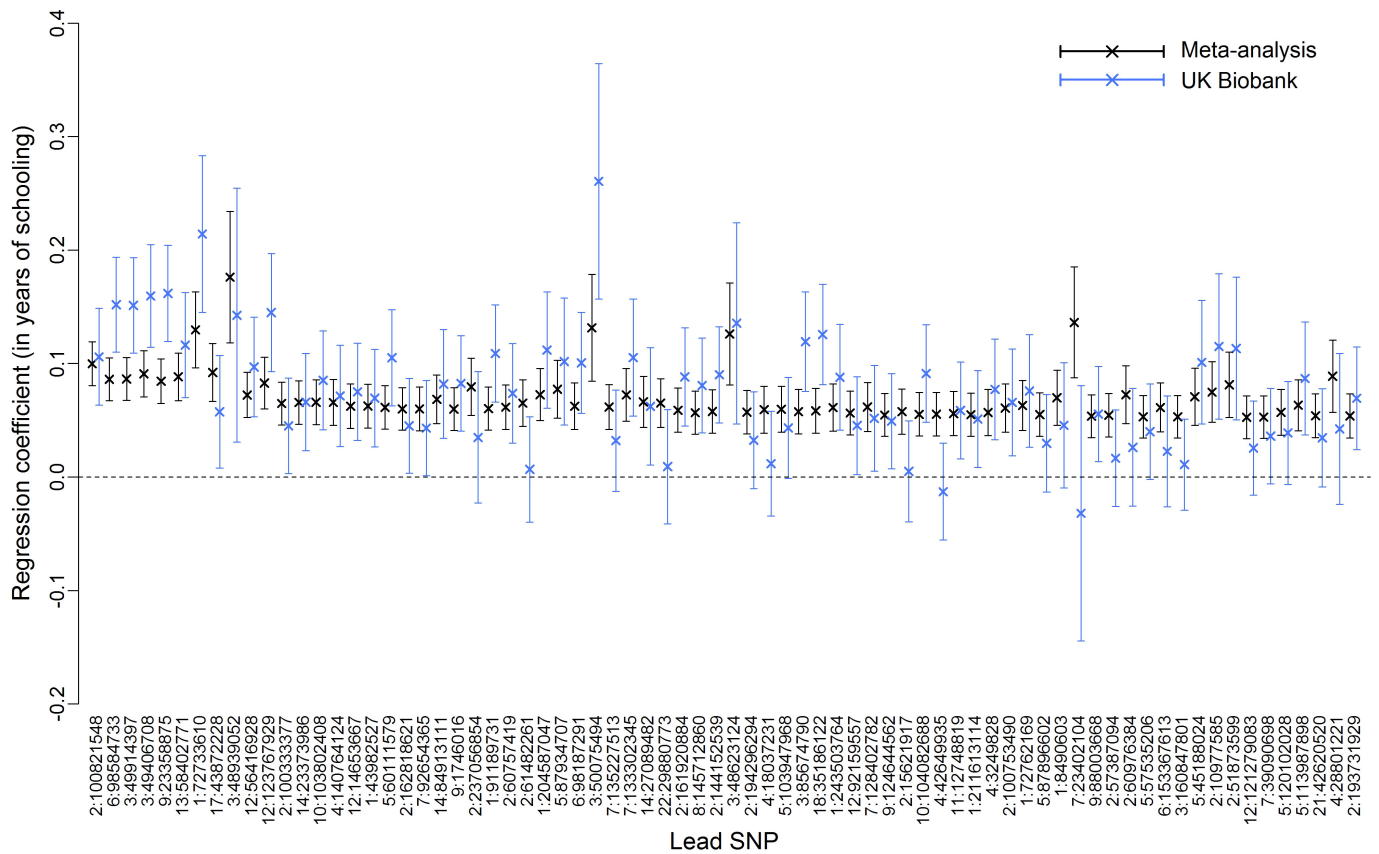
48 associations reported for waist-to-hip ratio adjusted for BMI (WHR). These results are based on the GIANT consortium's publicly available results for pooled analyses restricted to European-ancestry individuals: https://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium.



Extended Data Figure 3 | Assessing the extent to which population stratification affects the estimates from the GWAS. **a**, LD score regression plot with the summary statistics from the GWAS. Each point represents an LD score quantile for a chromosome (the x and y coordinates of the point are the mean LD score and the mean χ^2 statistic of variants in that quantile). That the intercept is close to 1 and that the χ^2 statistics increase linearly with the LD scores suggest that the bulk of the inflation in the χ^2 statistics is due to true polygenic signal and not to population

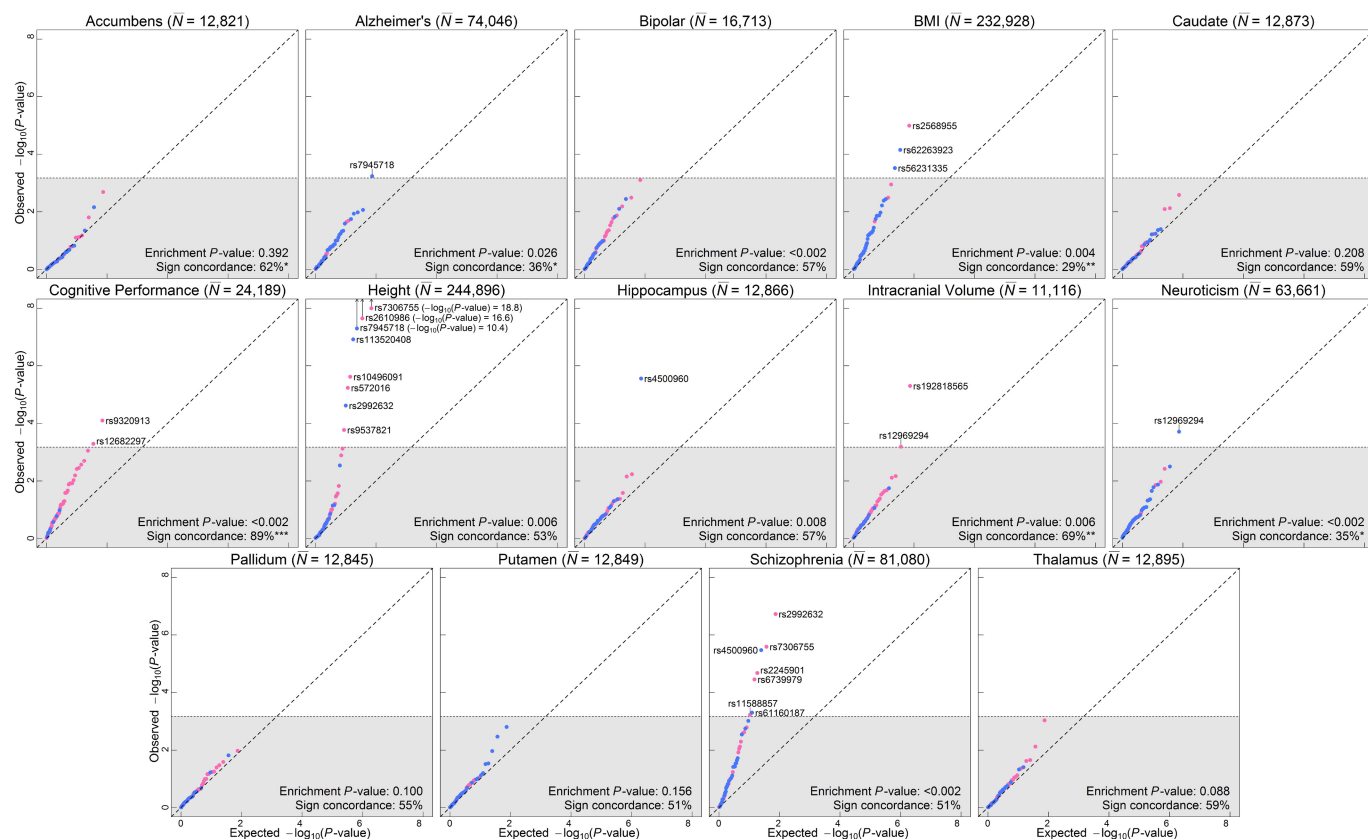


stratification. **b**, Estimates and 95% confidence intervals from individual-level and within-family regressions of *EduYears* on polygenic scores, for scores constructed with sets of SNPs meeting different P value thresholds. In addition to the analyses shown here, we conduct a sign concordance test, and we decompose the variance of the polygenic score. Overall, these analyses suggest that population stratification is unlikely to be a major concern for our 74 lead SNPs. See Supplementary Information section 3 for additional details.



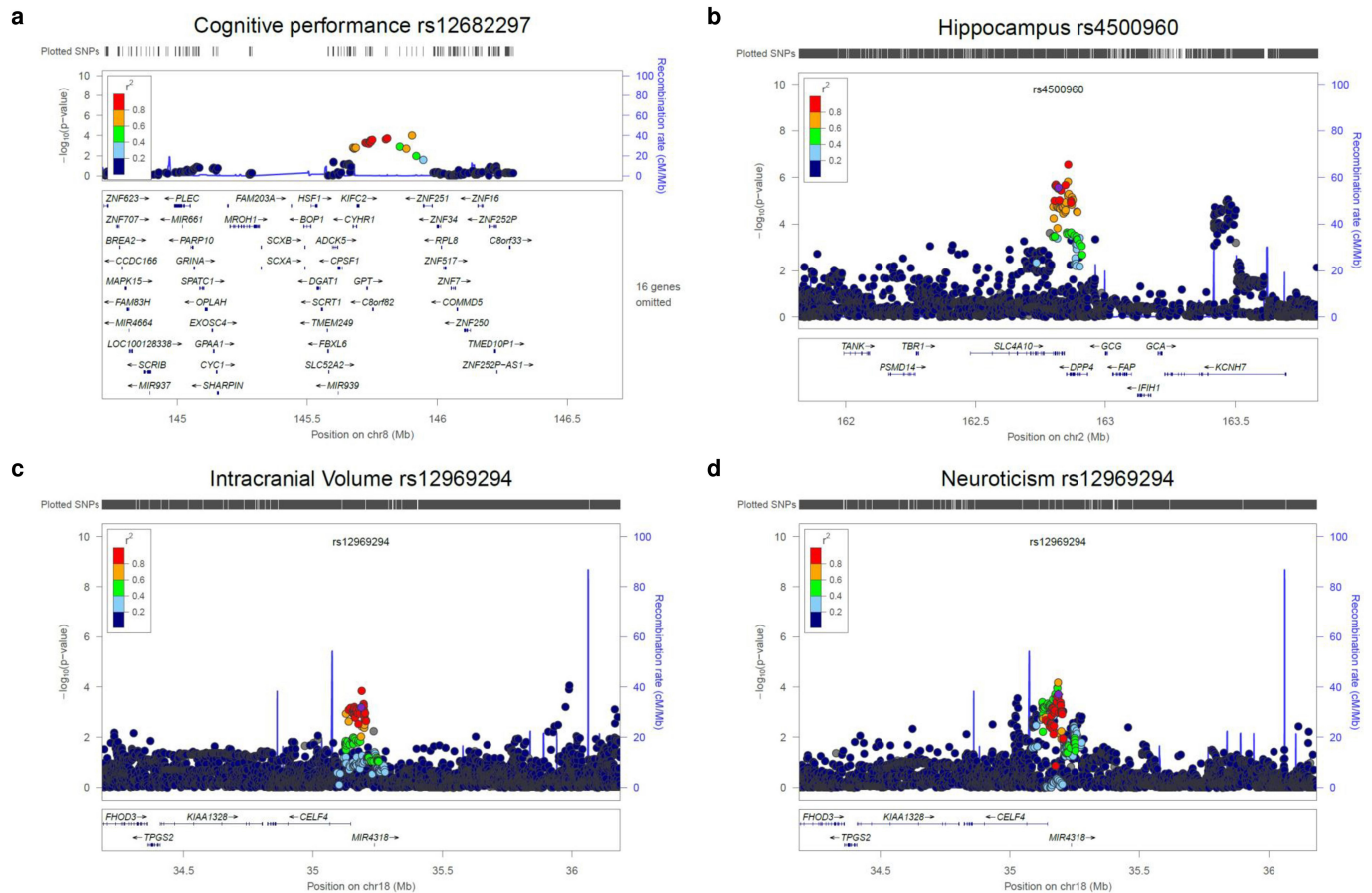
Extended Data Figure 4 | Replication of 74 lead SNPs in the UK Biobank data. Estimated effect sizes (in years of schooling) and 95% confidence intervals of the 74 lead SNPs in the meta-analysis sample ($n = 293,723$) and the UK Biobank replication sample ($n = 111,349$). The reference allele is the allele associated with higher values of EduYears

in the meta-analysis sample. SNPs are in descending order of R^2 in the meta-analysis sample. Of the 74 lead SNPs, 72 have the anticipated sign in the replication sample, 52 replicate at the 0.05 significance level, and 7 replicate at the 5×10^{-8} significance level.



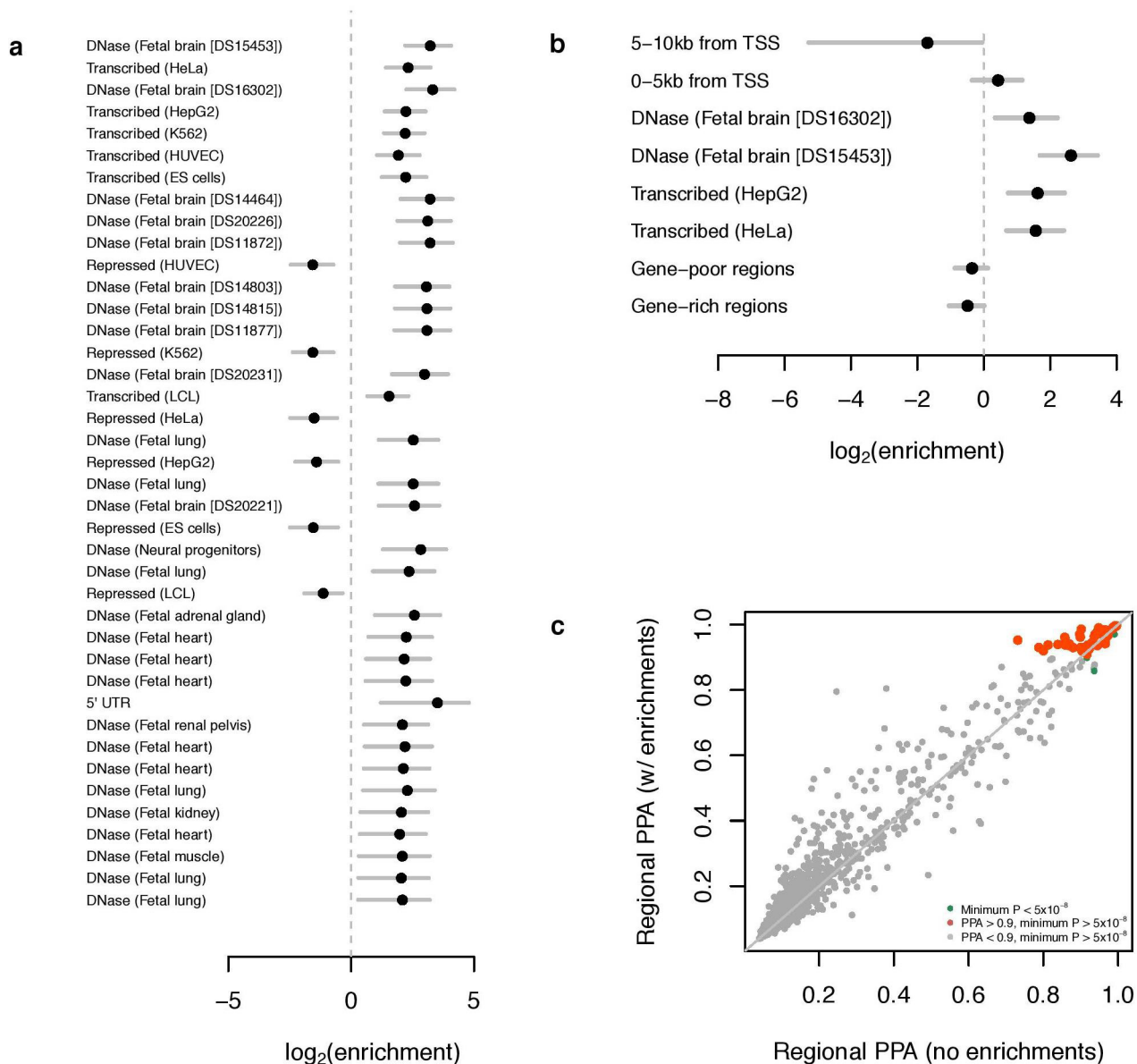
Extended Data Figure 5 | Q-Q plots for the 74 lead EduYears SNPs (or LD proxies) in published GWAS of other phenotypes. SNPs with concordant effects on both phenotypes are pink, and SNPs with discordant effects are blue. SNPs outside the grey area pass Bonferroni-corrected

significance thresholds that correct for the total number of SNPs we tested ($P < 0.05/74 = 6.8 \times 10^{-4}$) and are labelled with their rs numbers. Observed and expected P values are on a $-\log_{10}$ scale. For the sign concordance test: * $P < 0.05$, ** $P < 0.01$ and *** $P < 0.001$.



Extended Data Figure 6 | Regional association plots for four of the ten prioritized SNPs for mental health, brain anatomy, and anthropometric phenotypes identified using EduYears as a proxy phenotype. a, Cognitive performance; b, hippocampus; c, intracranial volume; d, neuroticism. The four were selected because very few

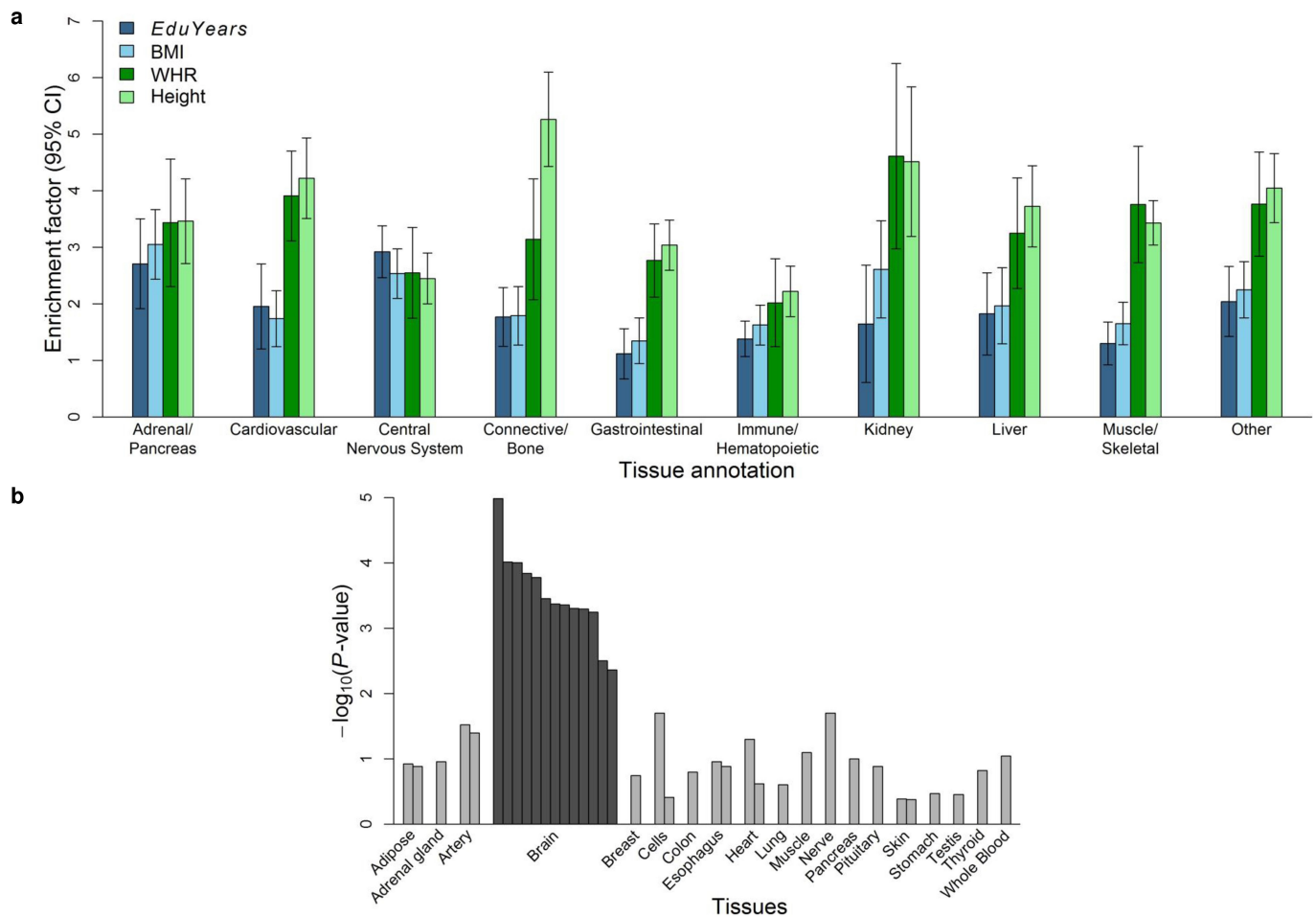
genome-wide significant SNPs have been previously reported for these traits. Data sources and methods are described in Supplementary Information section 3. The R^2 values are from the hg19 / 1000 Genomes Nov 2014 EUR references samples. The figures were created with LocusZoom (<http://csg.sph.umich.edu/locuszoom/>). Mb, megabases.



Extended Data Figure 7 | Application of fgwas to EduYears.

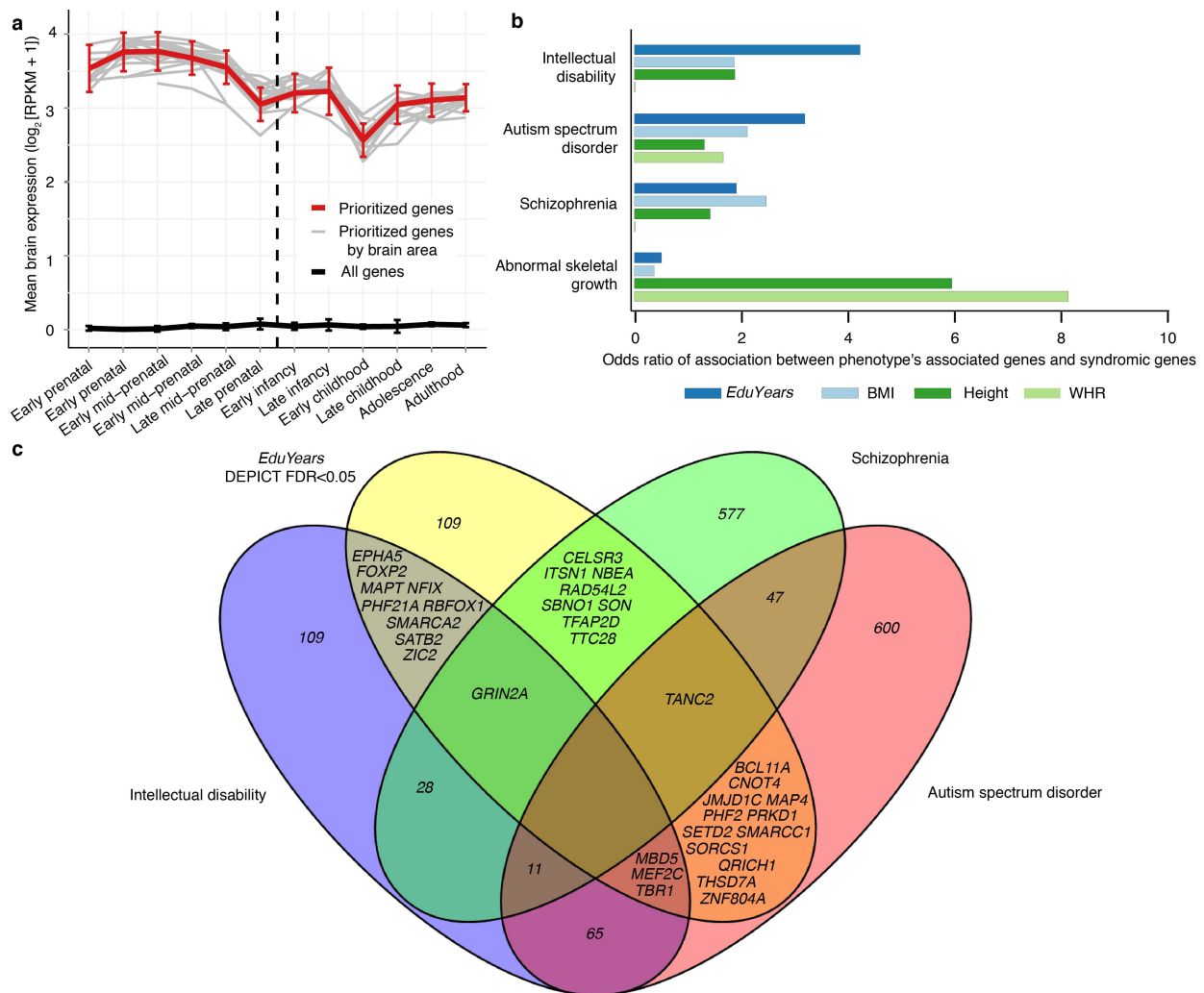
See Supplementary Information section 4.2 for further details. **a**, The results of single-annotation models. ‘Enrichment’ refers to the factor by which the prior odds of association at an LD-defined region must be multiplied if the region bears the given annotation; this factor is estimated using an empirical Bayes method applied to all SNPs in the GWAS meta-analysis regardless of statistical significance. Annotations were derived from ENCODE and a number of other data sources. Plotted are the base 2 logarithms of the enrichments and their 95% confidence intervals. Multiple instances of the same annotation correspond to independent replicates of the same experiment. **b**, The results of

combining multiple annotations and applying model selection and cross-validation. Although the maximum-likelihood estimates are plotted, model selection was performed with penalized likelihood. **c**, Reweighting of GWAS loci. Each point represents an LD-defined region of the genome, and shown are the regional posterior probabilities of association (PPAs). The x axis gives the PPA calculated from the GWAS summary statistics alone, whereas the y axis gives the PPA upon reweighting on the basis of the annotations in **b**. The orange points represent genomic regions where the PPA is equivalent to the standard GWAS significance threshold only upon reweighting.



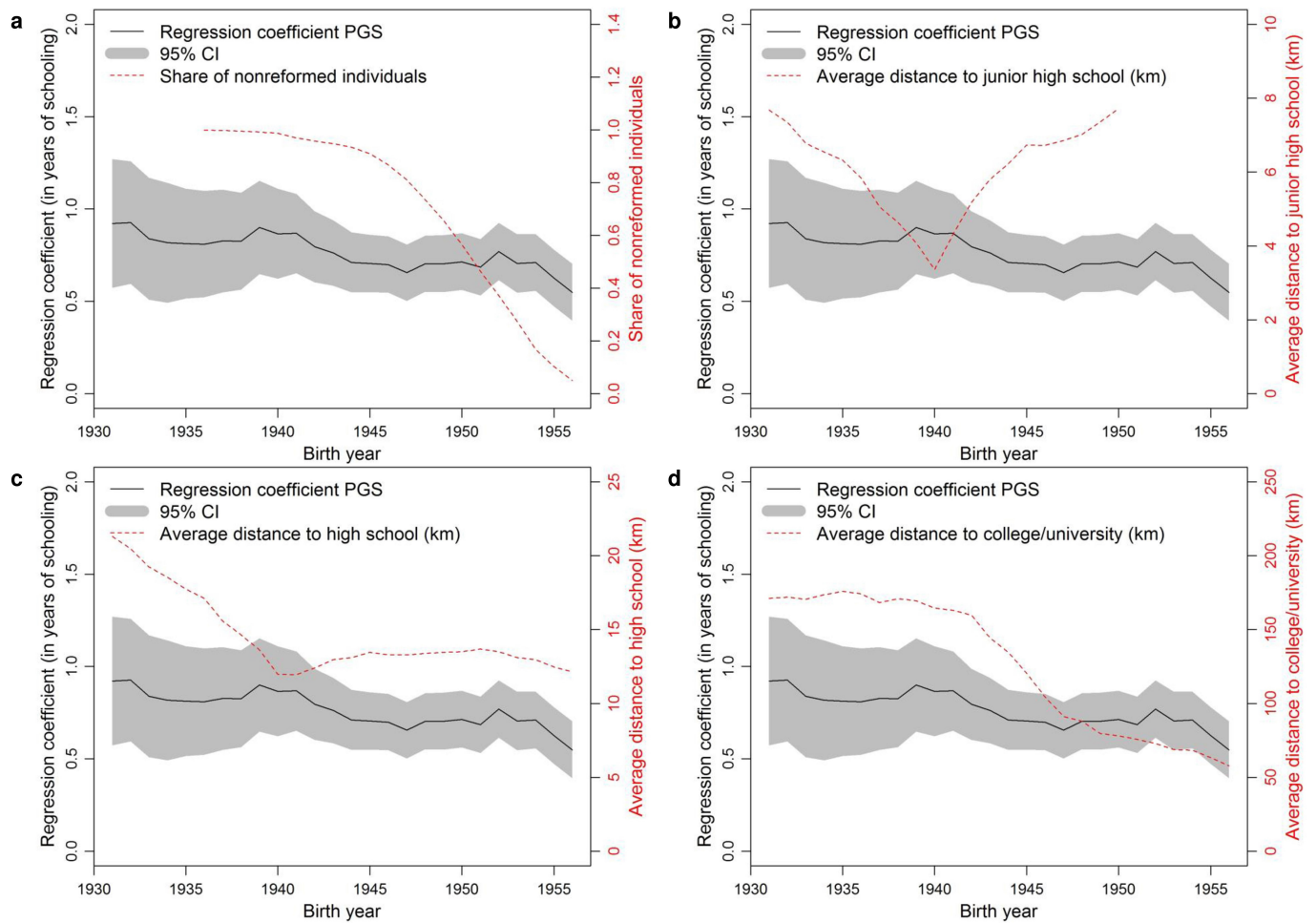
Extended Data Figure 8 | Tissue-level biological annotation. **a**, The enrichment factor for a given tissue type is the ratio of variance explained by SNPs in that group to the overall fraction of SNPs in that group. To benchmark the estimates for EduYears, we compare the enrichment factors to those obtained when we use the largest GWAS conducted to date on BMI, height, and waist-to-hip ratio adjusted for BMI. The estimates were produced with the LDSC Python software, using the LD scores and functional annotations introduced in ref. 17 and the HapMap3 SNPs with minor allele frequency >0.05 . Each of the ten enrichment calculations for a particular cell type is performed independently, while each controlling

for the 52 functional annotation categories in the full baseline model. The error bars show the 95% confidence intervals. **b**, We took measurements of gene expression by the Genotype-Tissue Expression (GTEx) Consortium and determined whether the genes overlapping EduYears-associated loci are significantly overexpressed (relative to genes in random sets of loci matched by gene density) in each of 37 tissue types. These types are grouped in the panel by organ. The dark bars correspond to tissues where there is significant overexpression. The y axis is the significance on a $-\log_{10}$ scale.



Extended Data Figure 9 | Gene-level biological annotation. **a**, The DEPict-prioritized genes for EduYears measured in the BrainSpan Developmental Transcriptome data (red curve) are more strongly expressed in the brain prenatally rather than postnatally. The DEPict-prioritized genes exhibit similar gene expression levels across different brain regions (grey lines). Analyses were based on \log_2 -transformed RNA-seq data. Error bars represent 95% confidence intervals. **b**, For

each phenotype and disorder, we calculated the overlap between the phenotype's DEPict-prioritized genes and genes believed to harbour *de novo* mutations causing the disorder. The bars correspond to odds ratios. **c**, DEPict-prioritized genes in EduYears-associated loci exhibit substantial overlap with genes previously reported to harbour sites where mutations increase risk of intellectual disability and autism spectrum disorder (Supplementary Table 4.6.1).



Extended Data Figure 10 | The predictive power of a polygenic score (PGS) varies in Sweden by birth cohort. Five-year rolling regressions of years of education on the PGS (left axis in all four panels), share of individuals not affected by the comprehensive school reform (**a**, right

axis), and average distance to nearest junior high school (**b**, right axis), nearest high school (**c**, right axis) and nearest college/university (**d**, right axis). The shaded area displays the 95% confidence intervals for the PGS effect.

Culturing of ‘unculturable’ human microbiota reveals novel taxa and extensive sporulation

Hilary P. Browne^{1*}, Samuel C. Forster^{1,2,3*}, Blessing O. Anonye¹, Nitin Kumar¹, B. Anne Neville¹, Mark D. Stares¹, David Goulding⁴ & Trevor D. Lawley¹

Our intestinal microbiota harbours a diverse bacterial community required for our health, sustenance and wellbeing^{1,2}. Intestinal colonization begins at birth and climaxes with the acquisition of two dominant groups of strict anaerobic bacteria belonging to the Firmicutes and Bacteroidetes phyla². Culture-independent, genomic approaches have transformed our understanding of the role of the human microbiome in health and many diseases¹. However, owing to the prevailing perception that our indigenous bacteria are largely recalcitrant to culture, many of their functions and phenotypes remain unknown³. Here we describe a novel workflow based on targeted phenotypic culturing linked to large-scale whole-genome sequencing, phylogenetic analysis and computational modelling that demonstrates that a substantial proportion of the intestinal bacteria are culturable. Applying this approach to healthy individuals, we isolated 137 bacterial species from characterized and candidate novel families, genera and species that were archived as pure cultures. Whole-genome and metagenomic sequencing, combined with computational and phenotypic analysis, suggests that at least 50–60% of the bacterial genera from the intestinal microbiota of a healthy individual produce resilient spores, specialized for host-to-host transmission. Our approach unlocks the human intestinal microbiota for phenotypic analysis and reveals how a marked proportion of oxygen-sensitive intestinal bacteria can be transmitted between individuals, affecting microbiota heritability.

A typical human intestinal microbiota contains 100–1,000 bacterial species with tremendous compositional diversity between individuals, such that each individual's microbiota is as unique as a fingerprint^{1,4}. Despite the taxonomic diversity, metagenomic sequencing has highlighted that a health-associated intestinal microbiome codes for highly conserved gene families and pathways associated with basic bacterial physiology and growth². However, many basic microbiota functions related to homeostasis, immune system development, digestion, pathogen resistance and microbiota inheritance have yet to be discovered⁵. This formidable challenge to validate and decipher the functional attributes of the microbiota has been hindered because the majority of intestinal bacteria are widely considered to be ‘unculturable’ and have never been isolated in the laboratory^{3,6}.

We sought to establish a genomic-based workflow that could be used as a platform for targeted culturing of specific bacterial phenotypes (Extended Data Fig. 1). Accordingly, we collected fresh faecal samples from six healthy humans and defined the resident bacterial communities with a combined metagenomic sequencing and bacterial culturing approach. Applying shotgun metagenomic sequencing, we profiled and compared the bacterial species present in the original faecal samples to those that grew as distinct colonies on agar plates containing the complex, broad-range bacteriological medium, YCFA⁷. Importantly, we observed a strong correlation between the two samples at the species level (Spearman's $\rho = 0.75$, $P < 0.01$) (Fig. 1a). When sequenced, the

original faecal sample and the cultured bacterial community shared an average of 93% of raw reads across the six donors. This overlap was 72% after *de novo* assembly (Extended Data Fig. 2). Comparison to a comprehensive gene catalogue that was derived by culture-independent means from the intestinal microbiota of 318 individuals⁴ found that 39.4% of the genes in the larger database were represented in our cohort and 73.5% of the 741 computationally derived metagenomic species identified through this analysis were also detectable in the cultured samples.

Together, these results demonstrate that a considerable proportion of the bacteria within the faecal microbiota can be cultured with a single growth medium. However, more than 8×10^6 distinct colonies would need to be picked from YCFA agar plates to match the species detection sensitivity of metagenomic sequencing. Thus, we established a broad-range culturing method that, when combined with high-throughput archiving or specific phenotypic selection, can be used to isolate and identify novel bacteria from the gastrointestinal tract.

The human intestinal microbiota is dominated by strict anaerobic bacteria that are extremely sensitive to ambient oxygen, so it is not known how these bacteria survive environmental exposure to be transmitted between individuals. Certain members of the Firmicutes phylum, including the diarrhoeal pathogen *Clostridium difficile*, produce metabolically dormant and highly resistant spores during colonization that facilitate both persistence within the host and environmental transmission^{8–10}. Relatively few intestinal spore-forming bacteria have been cultured to date, and while metagenomic studies suggest that other unexpected members of the intestinal microbiota possess potential sporulation genes, these bacteria remain poorly characterized^{11–14}.

We hypothesized that sporulation is an unappreciated basic phenotype of the human intestinal microbiota that may have a profound impact on microbiota persistence and spread between humans. Spores from *C. difficile* are resistant to ethanol and this phenotype can be used to select for spores from a mixed population of spores and ethanol-sensitive vegetative cells¹⁵. Faecal samples with or without ethanol treatment were processed using our combined culture and metagenomics workflow (Extended Data Fig. 1). Principle component analysis demonstrated that ethanol treatment profoundly altered the culturable bacterial composition and, when compared to the original profile, efficiently enriched for ethanol-resistant bacteria, facilitating their isolation (Fig. 1b). We picked ~2,000 individual bacterial colonies from both ethanol-treated and non-ethanol-treated conditions, re-streaked them to purity, and performed full-length 16S ribosomal RNA gene sequencing to enable taxonomic characterization. Unique taxa were then archived as frozen stocks for future phenotypic analysis.

In total, we archived bacteria representing 96% of the bacterial abundance at the genus level and 90% of the bacterial abundance at the species level based on average relative abundance across the six donors (Extended Data Fig. 3a, b). Even genera that were present at low average

¹Host-Microbiota Interactions Laboratory, Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK. ²Centre for Innate Immunity and Infectious Diseases, Hudson Institute of Medical Research, Clayton, Victoria 3168, Australia. ³Department of Molecular and Translational Sciences, Monash University, Clayton, Victoria 3800, Australia. ⁴Microbial Pathogenesis Laboratory, Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK.

*These authors contributed equally to this work.

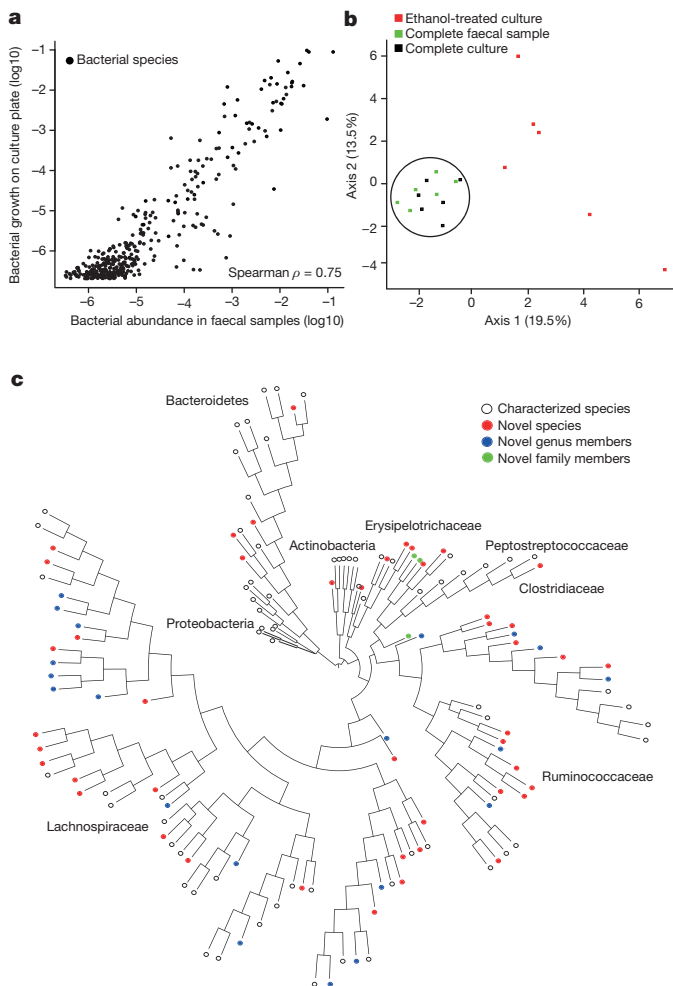


Figure 1 | Targeted phenotypic culturing facilitates bacterial discovery from healthy human faecal microbiota. **a**, Relative abundance of bacteria in faecal samples (x axis) compared with relative abundance of bacteria growing on YCFA agar plates (y axis) as determined by metagenomic sequencing. Bacteria grown on YCFA agar are representative of the complete faecal samples as indicated by Spearman $\rho = 0.75$ ($n = 6$). **b**, Principal component analysis plot of 16S rRNA gene sequences detected from six donor faecal samples ($n = 6$), representing bacteria in complete faecal samples (green), faecal bacterial colonies recovered from YCFA agar plates without ethanol pre-treatment (black) or with ethanol pre-treatment to select for ethanol-resistant spore-forming bacteria (red). Culturing without ethanol selection is representative of the complete faecal sample, ethanol treatment shifts the profile, enriching for ethanol-resistant spore-forming bacteria and allowing their subsequent isolation. **c**, Phylogenetic tree of bacteria cultured from the six donors constructed from full-length 16S rRNA gene sequences. Novel candidate species (red), genera (blue) and families (green) are shown by dot colours. Major phyla and family names are indicated. Proteobacteria were not cultured, but are included for context.

relative abundance ($<0.1\%$) were isolated (Extended Data Fig. 3c). Overall, we archived 137 distinct bacterial species including 45 candidate novel species (Fig. 1c, Extended Data Fig. 3d and Supplementary Table 1), and isolates representing 20 candidate novel genera and 2 candidate novel families. Our collection contains 90 species from the Human Microbiome Project's 'most wanted' list of previously uncultured and unsequenced microbes¹⁶ (Supplementary Table 1). Thus, our broad-range YCFA-based culturing approach led to massive bacterial discovery, and challenges the notion that the majority of the intestinal microbiota is unculturable.

We isolated and purified bacteria representing 66 distinct ethanol-resistant species that are distributed across 5 known families and

2 newly identified candidate families (Extended Data Fig. 3d and Extended Data Fig. 4). The identification of these new and unexpected spore-formers highlights the broad taxonomic distribution of this phenotype among the enteric species of the Firmicutes. To define the conserved genetic pathways underlying sporulation and germination within the intestinal microbiota, we sequenced, assembled and annotated the whole genomes of 234 archived ethanol-resistant and ethanol-sensitive bacteria. Previously, the gene markers used to identify spore-forming bacterial species have been based on underlying genetic assumptions^{13,17,18}; here we applied an unbiased computational approach to define 66 conserved genes linked to an ethanol-resistance phenotype (Extended Data Fig. 5 and Supplementary Table 1). This gene set allows for the prediction of the sporulation capabilities of bacterial species isolated from diverse environments with a high degree of accuracy (Extended Data Fig. 6a and Supplementary Table 1) and consists of genes from a wide range of functional classes (Extended Data Fig. 6b and Supplementary Table 1).

To test whether commensal spore formation facilitates long-term environmental survival, we exposed a phylogenetically diverse selection of commensal spore-forming and non-spore-forming bacteria and *C. difficile* to ambient oxygen for increasing periods of time. Under these conditions, non-spore-forming bacteria remained viable for 2–6 days (48–144 h) (Fig. 2a). In contrast, commensal spore-forming bacteria, *C. difficile* and the facultative anaerobe *Escherichia coli* were able to survive stably to the end of the experiment on day 21 (504 h). In addition, spore-forming commensals and *C. difficile*, but not non-spore-forming commensals, survived prolonged exposure to the common disinfectant ethanol (Extended Data Fig. 7). These results demonstrate that commensal spore-formers and *C. difficile* share a core set of sporulation genes that confer a highly resistant phenotype that is associated with environmental spread between humans.

C. difficile spores have evolved mechanisms to resume metabolism and vegetative growth after intestinal colonization by germinating in response to digestive bile acids released into the small intestine from the gall bladder⁹. We exposed enteric spore-formers and non-spore-formers to common bile acids (taurocholate, glycocholate and cholate) to assess their response to germinants after ethanol-shock treatment (Fig. 2b). Taurocholate was a potent germinant for all spore-formers, increasing the culturability of spores from commensal bacteria by between 8- and 70,000-fold ($P < 0.05$ for all spore-formers tested), whereas the other cholate derivatives had varying efficacy in germinating commensal spore-formers (Fig. 2b). Taurocholate and the other bile acids had no impact on the culturability of non-spore-formers, demonstrating that the effect is specific to spore-formers (Extended Data Fig. 8). We propose that this bile-acid-triggered 'colonizing germination' mechanism serves as a conserved *in vivo* cue to promote colonization by intestinal spore-forming bacteria. Thus, a duality of purpose exists in the *modus operandi* of intestinal spore-forming bacteria; spore formation ensures their survival and transmission while germination in response to *in vivo* cues ensures their persistence in the human population.

We next sought to estimate the proportions of spore-forming bacteria within the intestinal microbiota. Interrogation of the metagenomic data sets with the spore gene signature predicted that, on average, 60% of the genera contained spore-forming bacteria (Fig. 3a). These genera represent 30% of the total intestinal microbiota (Fig. 3b). We independently validated these observations with 16S rRNA gene amplicon sequencing (Extended Data Fig. 9). Importantly, these proportions of spore-forming bacteria were also observed in 1,351 publicly available faecal metagenomic data sets generated from healthy individuals¹⁹ (Fig. 3a, b). We also found the same proportion of spore-formers (61.3%) within the 'metagenomic species' derived from 318 healthy individuals⁴.

While the intestinal microbiota is considered to be relatively stable over time²⁰, evidence suggests that close contact of family members promotes sharing of Ruminococcaceae and Lachnospiraceae bacteria²¹, families that we describe as spore-formers (Extended Data Fig. 4).

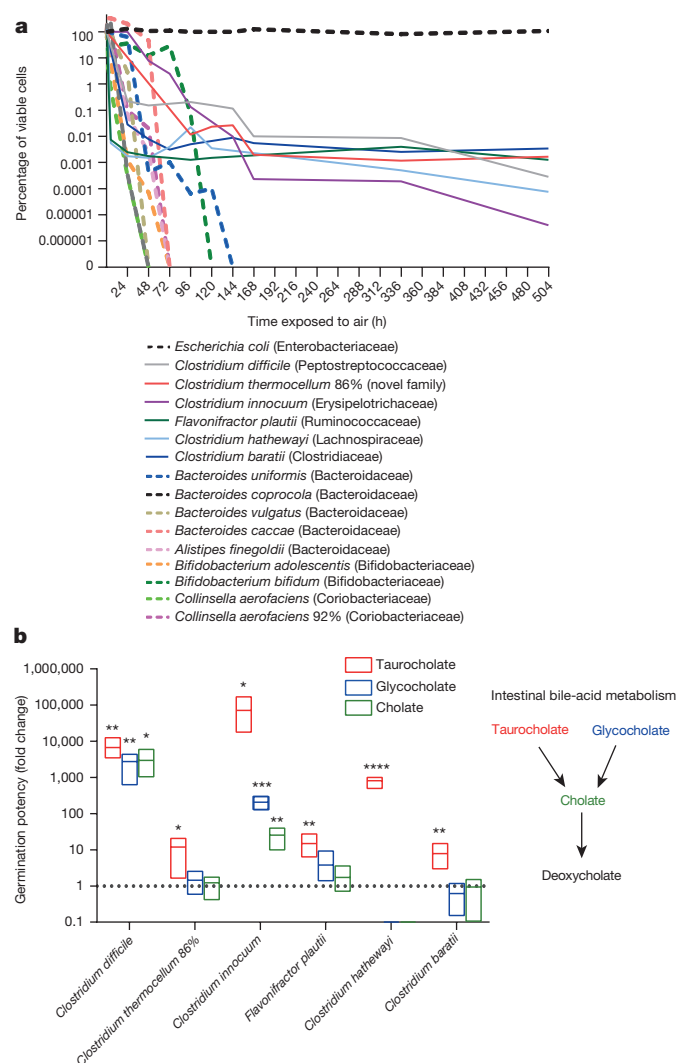


Figure 2 | Phenotypic characterization of phylogenetically diverse intestinal spore-forming bacteria. **a**, Spore-formers are more aero-tolerant than non-spore-formers, which is expected to facilitate host-to-host transmission. Once exposed to oxygen, only 1% of the original inoculum of non-spore-forming bacteria (dashed lines) were viable after 96 h (4 days) and none were viable after 144 h (6 days). Spore-forming bacteria (solid lines) persist owing to spore formation. The experiment was stopped after 504 h (21 days). Taxonomic families of each species tested are shown in brackets ($n = 3$ biological replicates for each strain). **b**, Intestinal spore-formers respond to bile-acid germinants. The number of colony-forming units (c.f.u.) (representing germinated spores) present on plates in the presence of a particular germinant is expressed as a fold change with respect to the number of c.f.u. recovered on plates in the absence of a germinant. Spore-formers and non-spore-formers were subjected to ethanol shock before being plated ($n = 6$ biological replicates for each strain). Only spore-formers survived. A fold change of one (dashed line) would indicate that a germinant had no effect on the number of c.f.u. recovered. Schematic summarizes the cholate-derived bile acid metabolism in the mammalian intestine. Mean and range, Welch's unpaired two-tailed t -test (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$).

We noted that in our cohort, the spore-forming bacteria of the microbiota were significantly more diverse than the non-spore-forming bacteria (Fig. 3c). To test the dynamics of the spore-forming and non-spore-forming bacteria over time, we analysed the metagenomic profiles of faecal samples collected from the same healthy subjects one year after the original sampling. Interestingly, we noted a significantly increased variability in the proportion of spore-forming bacteria compared with non-spore-forming bacteria over this period.

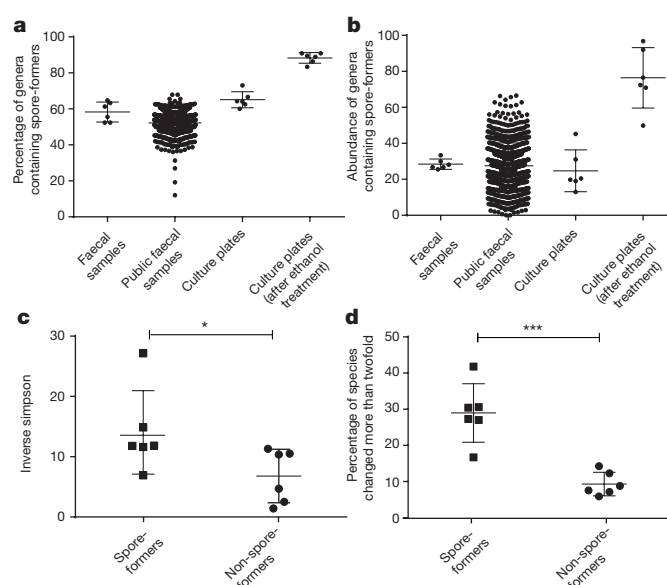


Figure 3 | Extensive and dynamic sporulation capacity within the human intestinal microbiota. **a**, **b**, Using the genomic signature to interrogate public ($n = 1,351$) and complete faecal sample metagenomic data sets from this study ($n = 6$) reveals the proportion of spore-formers as a count of the total number of genera (**a**) and as total microbial abundance (**b**). **c**, **d**, Metagenomic sequencing of donor faecal samples ($n = 6$) 1 year later demonstrates that spore-forming bacteria are more diverse than non-spore-forming bacteria (**c**) and that a significantly increased proportion of species show twofold or greater change over the same time period (**d**). Mean \pm standard deviation (s.d.), two-tailed paired t -test (* $P < 0.05$, *** $P < 0.001$).

This suggests a higher species turnover or a greater shift in relative abundance in the spore-forming bacterial species (Fig. 3d). Taken together, our phenotypic and genome analyses demonstrate that the spore-forming and non-spore-forming bacteria represent major, distinct phenotypic components of our microbiota, each with unique colonization dynamics.

We show that spore formation is a widespread, although previously unappreciated function of the human intestinal microbiota, with important implications for microbiota transmission and inheritance. On the basis of the shared phylogeny and common evolutionary and phenotypic characteristics of sporulation and germination, we propose that the abundant commensal intestinal spore-formers identified here rely on the same transmission and colonization strategy as *C. difficile*²². In brief, environmental *C. difficile* spores are highly transmissible for long periods after they are shed, commonly transmit within a local environment but also have the potential to spread rapidly over long distances²³. The transmission dynamics and geographical range of commensal spore-formers has yet to be determined, but we anticipate that this type of information will provide great insight into the heritability and the selective factors that shape the composition of the human intestinal microbiota.

Our workflow enables large-scale culturing, archiving, genome sequencing and phenotyping of novel bacteria from the human gut microbiota that were formerly considered to be unculturable. We have generated a sizable whole-genome-sequence data set that corresponds to 39% of the total number of intestinal bacterial genomes generated by the Human Microbiome Project. Our streamlined, single-medium approach, builds on the considerable efforts of others^{24,25} and unlocks the human intestinal microbiota for phenotypic characterization.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 25 September 2015; accepted 8 March 2016.

Published online 4 May 2016.

- Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
- Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220–230 (2012).
- Walker, A. W., Duncan, S. H., Louis, P. & Flint, H. J. Phylogeny, culturing, and metagenomics of the human gut microbiota. *Trends Microbiol.* **22**, 267–274 (2014).
- Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnol.* **32**, 822–828 (2014).
- Nelson, K. E. *et al.* A catalog of reference genomes from the human microbiome. *Science* **328**, 994–999 (2010).
- Stewart, E. J. Growing unculturable bacteria. *J. Bacteriol.* **194**, 4151–4160 (2012).
- Duncan, S. H., Hold, G. L., Harmsen, H. J., Stewart, C. S. & Flint, H. J. Growth requirements and fermentation products of *Fusobacterium prausnitzii*, and a proposal to reclassify it as *Faecalibacterium prausnitzii* gen. nov., comb. nov. *Int. J. Syst. Evol. Microbiol.* **52**, 2141–2146 (2002).
- Lawley, T. D. *et al.* Antibiotic treatment of *Clostridium difficile* carrier mice triggers a supershedder state, spore-mediated transmission, and severe disease in immunocompromised hosts. *Infect. Immun.* **77**, 3661–3669 (2009).
- Francis, M. B., Allen, C. A., Shrestha, R. & Sorg, J. A. Bile acid recognition by the *Clostridium difficile* germinant receptor, CspC, is important for establishing infection. *PLoS Pathog.* **9**, e1003356 (2013).
- Janoir, C. *et al.* Adaptive strategies and pathogenesis of *Clostridium difficile* from *in vivo* transcriptomics. *Infect. Immun.* **81**, 3757–3769 (2013).
- Rajilić-Stojanović, M. & de Vos, W. M. The first 1000 cultured species of the human gastrointestinal microbiota. *FEMS Microbiol. Rev.* **38**, 996–1047 (2014).
- Galperin, M. Y. *et al.* Genomic determinants of sporulation in *Bacilli* and *Clostridia*: towards the minimal set of sporulation-specific genes. *Environ. Microbiol.* **14**, 2870–2890 (2012).
- Abecasis, A. B. *et al.* A genomic signature and the identification of new sporulation genes. *J. Bacteriol.* **195**, 2101–2115 (2013).
- Meehan, C. J. & Beiko, R. G. A phylogenomic view of ecological specialization in the *Lachnospiraceae*, a family of digestive tract-associated bacteria. *Genome Biol. Evol.* **6**, 703–713 (2014).
- Riley, T. V., Brazier, J. S., Hassan, H., Williams, K. & Phillips, K. D. Comparison of alcohol shock enrichment and selective enrichment for the isolation of *Clostridium difficile*. *Epidemiol. Infect.* **99**, 355–359 (1987).
- Fodor, A. A. *et al.* The “most wanted” taxa from the human microbiome for whole genome sequencing. *PLoS One* **7**, e41294 (2012).
- de Hoon, M. J., Eichenberger, P. & Vitkup, D. Hierarchical evolution of the bacterial sporulation network. *Curr. Biol.* **20**, R735–R745 (2010).
- Onyenwoke, R. U., Brill, J. A., Farahi, K. & Wiegell, J. Sporulation genes in members of the low G+C Gram-type-positive phylogenetic branch (Firmicutes). *Arch. Microbiol.* **182**, 182–192 (2004).
- Forster, S. C. *et al.* HPMCD: the database of human microbial communities from metagenomic datasets and microbial reference genomes. *Nucleic Acids Res.* **44**, D604–D609 (2016).
- Faith, J. J. *et al.* The long-term stability of the human gut microbiota. *Science* **341**, 1237439 (2013).
- Schloss, P. D., Iverson, K. D., Petrosino, J. F. & Schloss, S. J. The dynamics of a family's gut microbiota reveal variations on a theme. *Microbiome* **2**, 25 (2014).
- Paredes-Sabja, D., Shen, A. & Sorg, J. A. *Clostridium difficile* spore biology: sporulation, germination, and spore structural proteins. *Trends Microbiol.* **22**, 406–416 (2014).
- He, M. *et al.* Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. *Nature Genet.* **45**, 109–113 (2013).
- Goodman, A. L. *et al.* Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proc. Natl Acad. Sci. USA* **108**, 6252–6257 (2011).
- Lagier, J. C. *et al.* Microbial culturomics: paradigm shift in the human gut microbiome study. *Clin. Microbiol. Infect.* **18**, 1185–1193 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the Wellcome Trust (098051); the United Kingdom Medical Research Council (PF451 to T.D.L.); the Australian National Health and Medical Research Council (1091097 to S.C.F.) and the Victorian Government's Operational Infrastructure Support Program (S.C.F.). We are grateful to G. Dougan and A. Walker for their input. We would also like to acknowledge funding from the Wellcome Trust Sanger Institute Technology Translation team, and sequencing and bioinformatics support from the Pathogen Informatics team.

Author Contributions H.P.B., B.O.A. and M.D.S. developed culturing procedures; H.P.B. carried out anaerobic culturing and bacterial isolation; D.G. prepared TEM images; S.C.F., N.K. and H.P.B. performed bioinformatics analyses; H.P.B., S.C.F. and T.D.L. designed the study. H.P.B., S.C.F., B.A.N. and T.D.L. analysed data and wrote the paper.

Author Information Assembled and annotated genome sequence data have been deposited in the European Nucleotide Archive under accession number ERP012217. Bacterial isolates have been deposited at the Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures (<http://www.dsmz.de>), the CCUG-Culture Collection, University of Gothenburg, Sweden (<http://www.ccug.se>), the Belgian Co-ordinated Collection of Micro-organisms hosted by the Laboratory of Microbiology (BCCM/LMG) at Ghent University (<http://bccm.belspo.be/>) and at the Japan Collection of Microorganisms (JCM; <http://jcm.brc.riken.jp/en/>). Isolate accession numbers are listed in Supplementary Table 1. Any isolates without accession numbers are available upon request. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.D.L. (tl2@sanger.ac.uk).



This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

METHODS

Culturing. Fresh faecal samples were obtained from six consenting healthy adult human donors (1 faecal sample per donor: minimum 0.5 g) and were placed in anaerobic conditions within 1 h of passing to preserve the viability of anaerobic bacteria. All sample processing and culturing took place under anaerobic conditions in a Whitley DG250 workstation at 37°C. Culture media, PBS and all other materials that were used for culturing were placed in the anaerobic cabinet 24 h before use to reduce to anaerobic conditions. The faecal samples were divided in two. One part was homogenized in reduced PBS (0.1 g stool per ml PBS) and was serially diluted and plated directly onto YCFA⁷ agar supplemented with 0.002 g ml⁻¹ each of glucose, maltose and cellobiose in large (13.5 cm diameter) Petri dishes. This sample was also subjected to metagenomic sequencing to profile the entire community. The other part was treated with an equal volume of 70% (v/v) ethanol for 4 h at room temperature under ambient aerobic conditions to kill vegetative cells. Then, the solid material was washed three times with PBS and it was eventually resuspended in PBS. Plating was performed as described earlier.

For the ethanol-treated samples, the medium was supplemented with 0.1% sodium taurocholate to stimulate spore germination. Colonies were picked 72 h after plating from Petri dishes of both ethanol-treated and non-ethanol-treated conditions harbouring non-confluent growth, (that is, plates on which the colonies were distinct and not touching). The colonies that were picked were re-streaked to confirm purity. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Microbiota profiling and sequencing. Identification of each isolate was performed by PCR amplification of the full-length 16S rRNA gene (using 7F (5'-AGAGTTTGTATYMTGGCTCAG-3') forward primer and 1510R (5'-ACGGYTACCTTGTACGACTT-3') reverse primer followed by capillary sequencing. Full-length 16S rRNA gene sequence reads were aligned in the Ribosomal Database Project (RDP), manually curated in ARB²⁶ and mothur²⁷ was then used to classify reads to operational taxonomic units (OTUs). The R package seqinr version 3.1 was used to determine sequence similarity between OTUs and 98.7% was used as a species-level cut-off^{28,29}. The full-length 16S rRNA gene sequence of each species-level OTU was compared to the RDP reference database to assign taxonomic designations to the genus level³⁰ and a BLASTn search defined either a characterized or candidate novel species³¹.

Comparisons with the Human Microbiome Project (HMP) were carried out using 97% sequence similarity of the 16S rRNA gene sequence from the cultured bacteria to define a species because only partial 16S rRNA gene sequences were available. HMP data regarding the most wanted taxa and the completed sequencing projects were downloaded from http://hmpdacc.org/most_wanted/#data and <http://hmpdacc.org/HMRGD/>, respectively.

Genomic DNA was extracted from at least one representative of each unique OTU using a phenol-chloroform-based DNA isolation procedure. DNA was sequenced on the Illumina HiSeq platform generating read lengths of 100 bp and these were assembled and annotated for further analysis. DNA was extracted directly from each faecal sample for whole-community metagenomic and 16S rRNA gene amplicon sequencing using the MP Biomedical FastDNA SPIN Kit for soil. To enable comparisons with the complete community samples, non-confluent cultures were scraped from agar plates 72 h after inoculation with the initial faecal sample and DNA was extracted from this community using the same DNA isolation process. 16S rRNA gene amplicon libraries were made by PCR amplification of variable regions 1 and 2 of the 16S rRNA gene using the Q5 High-Fidelity Polymerase Kit supplied by New England Biolabs. Primers 27F AATGATACGGCGACCACCGAGATCTACAC (first part, Illumina adaptor) TATGTTAATT (second part, forward primer pad) CC (third part, forward primer linker) AGMGTTYGATYMTGGCTCAG (fourth part, forward primer) and 338R CAAGCAGAAGACGGCATACGAGAT (first part, reverse complement of 3' Illumina adaptor) ACGAGACTGATT (second part, golay barcode) AGTCAGTCAG (third part, reverse primer pad) AA (fourth part, reverse primer linker) GCTGCTCCCGTAGGAGT (fifth part, reverse primer) were used. Four PCR amplification reactions per sample were carried out; products were pooled and combined in equimolar amounts for sequencing using the Illumina MiSeq platform, generating 150 bp reads.

Microbiota analysis. A maximum likelihood phylogeny of the culture-derived bacteria was generated from the aligned RDP sequence using FastTree version 2.1.3 (ref. 32) with the following settings: a generalized time-reversible (GTR) model of nucleotide substitution and CAT approximation of the variation in rates across sites with 20 rate categories. The ethanol-resistant phylogeny was derived directly from the entire culture phylogeny. All phylogenetic trees were edited in ITOL³³.

Analysis of the partial 16S rRNA gene sequence generated from the 16S rRNA gene amplicon libraries was carried out using the mothur MiSeq SOP³⁴ on

29 August 2014, generating 7,549 OTUs across all samples. A sequence similarity threshold of 97% was used to define an OTU.

Metagenomic sequence reads were analysed using the Kraken³⁵ taxonomic sequence classification approach based on a custom database comprising complete, high-quality reference bacterial, DNA viral and archaeal genomes in addition to the genomes sequenced in this research. Resulting classified reads were log₂ transformed and standardized by total abundance. Metagenomic samples were compared at the genus and species levels by relative abundance and at the genetic level by alignment using the bowtie2 algorithm³⁶ to the appropriate gene catalogue. Sequences were considered present where an average of twofold coverage was achieved across the length of the considered sequence. A cut-off of 100 unique reads was applied to determine metagenomic species detection. Where appropriate, Spearman's rank correlation coefficient was applied for correlation analysis. Inverse Simpson's diversity index was calculated from Kraken output in R version 3.2.1 using the vegan: Community Ecology Package version 2.3-0.

Gene sporulation signature. Heuristic based bidirectional best hit analysis was performed to identify 21,342 conserved genes within the 694,300 genes annotated across the 234 sequenced genomes. Support vector, machine-based, contrast set association mining was applied to identify the optimal, weighted gene signature consisting of 66 genes. Species classification was performed using BLAST-based gene detection with percentage detection weighted by gene signature contribution and scaled to generate a total score between 0 and 1. Scores greater than 0.5 were considered true spore-formers based on comparison to known spore-formers. Signature-based abundance was assessed against 1,351 publically available metagenomic data sets from healthy individuals¹⁹ after taxonomic assignment using the Kraken database. Genera were considered spore-formers when all known species within that genus had a spore forming score greater than 0.5.

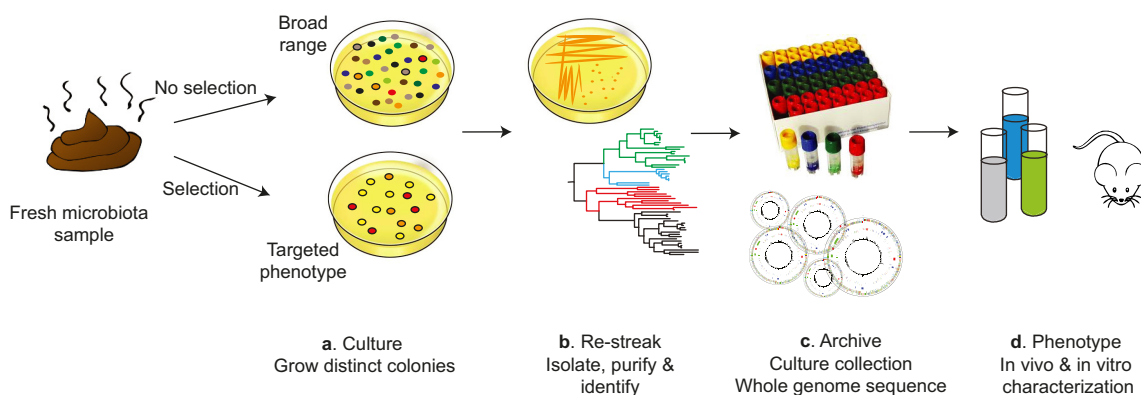
Transmission electron microscopy. Spore images were generated using transmission electron microscopy (TEM) as previously described³⁷. Bacterial isolates for imaging were prepared by streaking pure cultures from frozen glycerol stocks and confirming purity by full-length 16S rRNA gene sequencing after one round of sub-culture to obtain visible and isolated single colonies. TEM images were prepared from culture plates 72 h after inoculation. The number of spore bodies visible in the TEM images was expressed as a percentage of the number of vegetative cells present and this ranged from 1% for *Ruminococcus flavefaciens*_93% to 4% for *Turicibacter sanguinis*.

Oxygen sensitivity assay. Pure cultures were grown overnight in YCFA broth under anaerobic culture conditions as described earlier and the cultures were spotted in a dilution series onto YCFA agar containing 0.1% sodium taurocholate. Plates were incubated under ambient (aerobic) conditions at room temperature for specified time periods before being returned to the anaerobic cabinet. Colony-forming units (c.f.u.) were counted 72 h later. Cultures that were incubated anaerobically, and which were therefore not exposed to oxygen, acted as controls. Prior to the assay, all species were subjected to ethanol shock and were cultured anaerobically to determine their ability to sporulate. The viability of the oxygen-exposed cultures was expressed as a percentage of the viability of the anaerobic control cultures.

Germination response to intestinal bile acids assay. Pure cultures were grown overnight in YCFA broth under anaerobic conditions and were then washed by repeatedly centrifuging to a pellet and re-suspending in PBS. Vegetative cells were killed using an ethanol shock treatment as previously described and the cultures were then serially diluted and plated on YCFA agar with and without 0.1% intestinal bile salts (taurocholate, cholate and glycocholate). Colony-forming units (c.f.u.) were counted 72 h later and the fold change of the number of c.f.u. present on plates in the presence of a particular germinant with respect to the number of c.f.u. present on plates in the absence of a germinant was calculated. The limit of detection (200 c.f.u. ml⁻¹) was used for the number of c.f.u. recovered from *Clostridium hathewayi* plated without any germinants to allow a fold-change calculation. The experiment to determine the response of non-spore-formers to germinants was carried out similarly, except that vegetative cells were not treated with ethanol but rather were serially diluted and plated directly after washing.

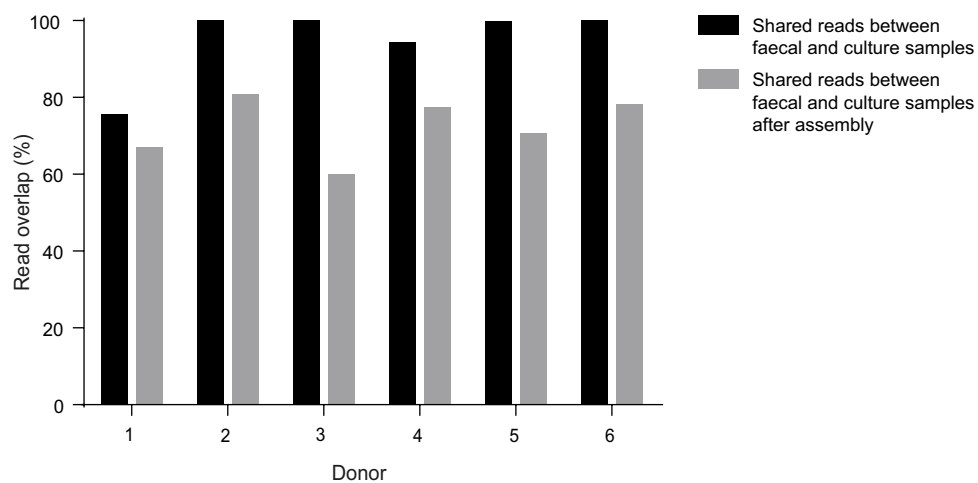
26. Ludwig, W. *et al.* ARB: a software environment for sequence data. *Nucleic Acids Res.* **32**, 1363–1371 (2004).
27. Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**, 7537–7541 (2009).
28. Bosshard, P. P., Abels, S., Zbinden, R., Böttger, E. C. & Altwegg, M. Ribosomal DNA sequencing for identification of aerobic gram-positive rods in the clinical laboratory (an 18-month evaluation). *J. Clin. Microbiol.* **41**, 4134–4140 (2003).
29. Clarridge, J. E. III. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin. Microbiol. Rev.* **17**, 840–862 (2004).

30. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007).
31. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
32. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
33. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39**, W475–W478 (2011).
34. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.* **79**, 5112–5120 (2013).
35. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
36. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).
37. Lawley, T. D. et al. Proteomic and genomic characterization of highly infectious *Clostridium difficile* 630 spores. *J. Bacteriol.* **191**, 5377–5386 (2009).
38. Bosshard, P. P., Zbinden, R. & Altwegg, M. *Turicibacter sanguinis* gen. nov., sp. nov., a novel anaerobic, Gram-positive bacterium. *Int. J. Syst. Evol. Microbiol.* **52**, 1263–1266 (2002).
39. Duncan, S. H., Hold, G. L., Barcenilla, A., Stewart, C. S. & Flint, H. J. *Roseburia intestinalis* sp. nov., a novel saccharolytic, butyrate-producing bacterium from human faeces. *Int. J. Syst. Evol. Microbiol.* **52**, 1615–1620 (2002).
40. Iino, T., Mori, K., Tanaka, K., Suzuki, K. & Harayama, S. *Oscillibacter valericigenes* gen. nov., sp. nov., a valerate-producing anaerobic bacterium isolated from the alimentary canal of a Japanese corbicula clam. *Int. J. Syst. Evol. Microbiol.* **57**, 1840–1845 (2007).
41. Paredes-Sabja, D., Setlow, P. & Sarker, M. R. Germination of spores of *Bacillales* and *Clostridiales* species: mechanisms and proteins involved. *Trends Microbiol.* **19**, 85–94 (2011).



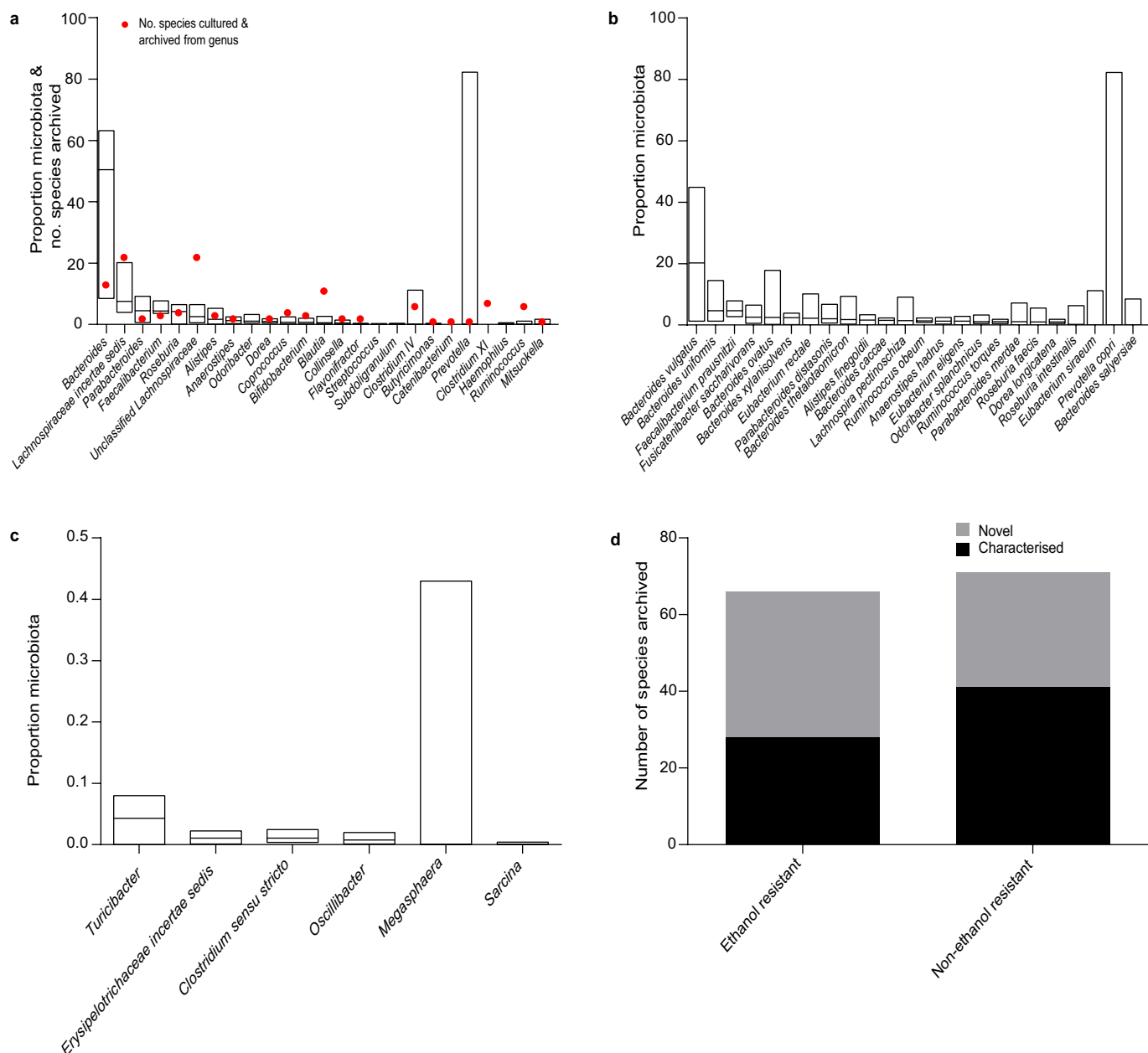
Extended Data Figure 1 | A workflow for culturing, archiving and characterization of the intestinal microbiota. a–d, Schematic diagram of the workflow, encompassing bacterial culturing and genomics to isolate and characterize bacterial species from the human intestinal microbiota. The process incorporates several steps, which are culture, re-streak, archive and phenotype. **a,** Fresh faecal samples are left untreated or are treated to select for bacteria with a desired phenotype (such as sporulation). The stool is homogenized and then serially diluted and

then aliquots of the homogenate are inoculated on YCFA agar to culture bacteria. **b,** Isolates are identified by selecting single colonies that are streaked to purity and full-length 16S rRNA genes are amplified and sequenced. **c,** Each unique, novel and desired isolate is archived frozen in a culture collection and a whole-genome sequence is generated for each. **d,** Phenotypic characterization and functional validation of metagenomics studies can be performed *in vitro* and *in vivo*.



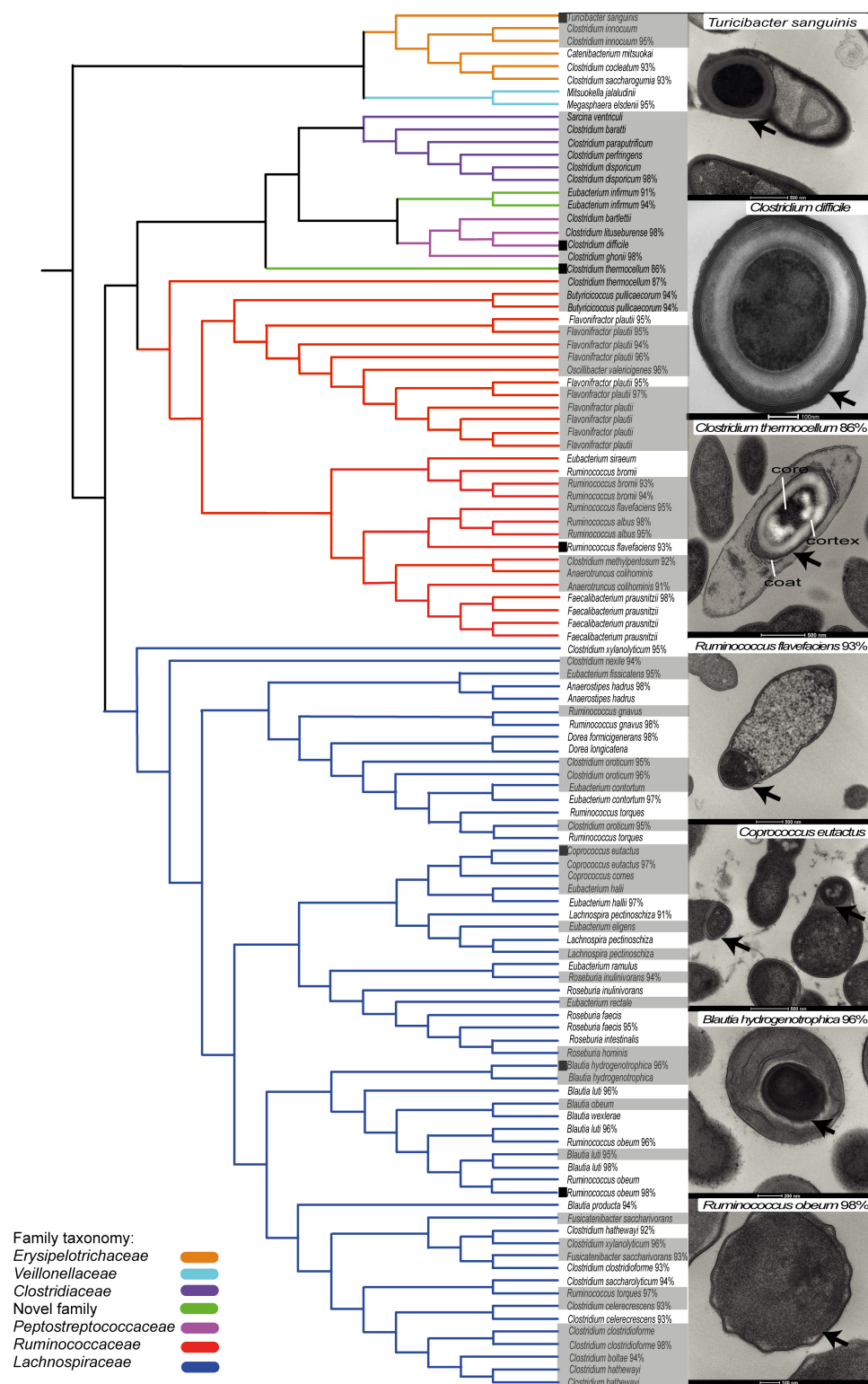
Extended Data Figure 2 | Comparison of sequence read content of faecal samples and cultured samples for six donors. The majority of sequence reads from the original donor faecal samples ($n = 6$) are present

in culture samples both as raw reads (93% shared on average across the six donors) and after *de novo* assembly (72% shared on average across the six donors).



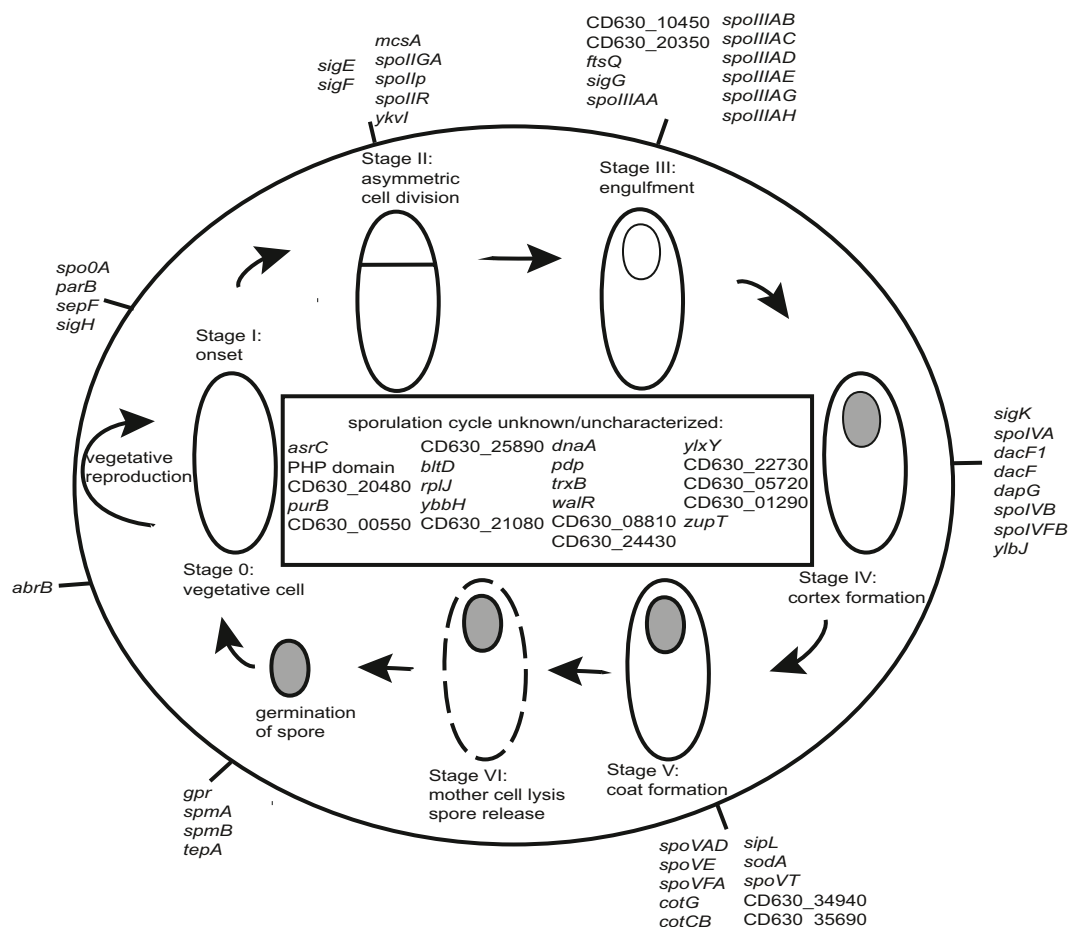
Extended Data Figure 3 | Archiving of bacterial diversity and novelty through anaerobic culturing. **a**, **b**, Representative species from 21 of the 25 most abundant bacterial genera (**a**) and 23 of the 24 most abundant species (**b**) were isolated and archived (abundance was determined by metagenomic sequencing and based on average relative abundance across the six donors ($n = 6$)). This represents 96% of the average relative abundance at the genus level and 90% of the average relative abundance at the species level across the six donors. A red dot in **a** indicates the number of species archived from each genus. *Lachnospiraceae incertae sedis*,

unclassified *Lachnospiraceae*, *Clostridium IV* and *Clostridium XI* are not strict genera and represent currently unclassified species. *Odoribacter splanchnicus* in **b** was the only species not archived. **c**, Lowly represented intestinal microbiota members were also cultured. At least one representative species from each of the genera presented were cultured. Median and range is presented for the above with taxa ranked by median value. **d**, The number of bacterial species cultured in this study. At least 40% from each category were previously unknown.



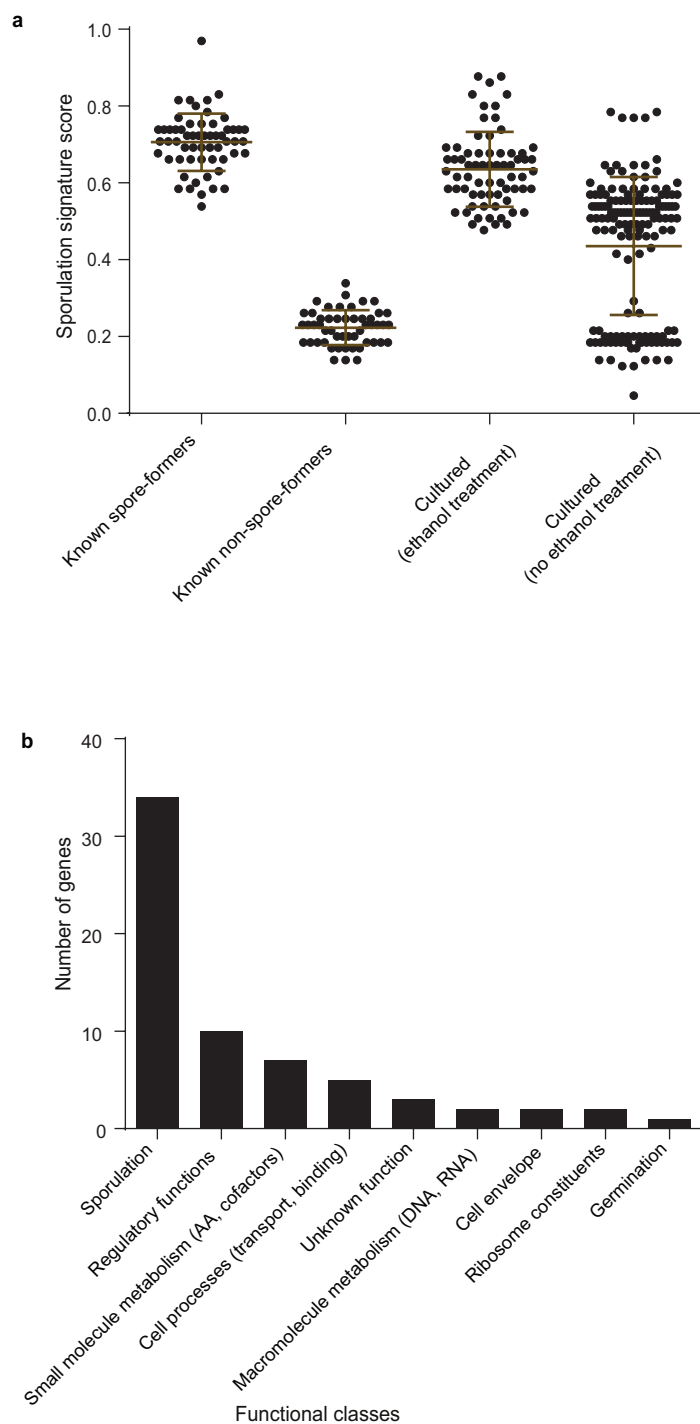
Extended Data Figure 4 | Phylogeny of intestinal spore-forming bacteria. Full length 16S rRNA gene phylogeny illustrating the taxonomic relationship of ethanol-resistant bacteria within the Firmicutes cultured from the donor faecal samples. Branch colours indicate distinct families. Shaded text indicates species cultured from an ethanol-treated faecal sample and unshaded text indicates species cultured from a non-ethanol-treated faecal sample. Percentage values represent closest identity to a characterized species. Transmission electron micrographs (TEMs) of spore ultrastructures for a phylogenetically diverse selection of cultured bacteria are shown with an arrow in images and include a candidate novel family with 86% identity to the 16S rRNA gene sequence from *Clostridium*

thermocellum. Typical spore structures are defined and illustrated in the same image. TEMs are ordered according to boxes next to the species name. Scale bars are shown at the bottom of each image. *C. difficile* is included for context. Bacteria displaying an ethanol-resistant phenotype represent species previously classified as non-spore-formers (*Turicibacter sanguinis*³⁸ and closely related candidate novel species), species closely related to non-spore-formers (*Roseburia intestinalis*³⁹ and *Oscillibacter valericigenes*⁴⁰ and closely related candidate novel species) or species suspected of forming spores but which, to our knowledge, have never been demonstrated to do so until now (*Eubacterium eligens*, *Eubacterium rectale*, *Coproccoccus comes*⁴¹ and related candidate novel species).



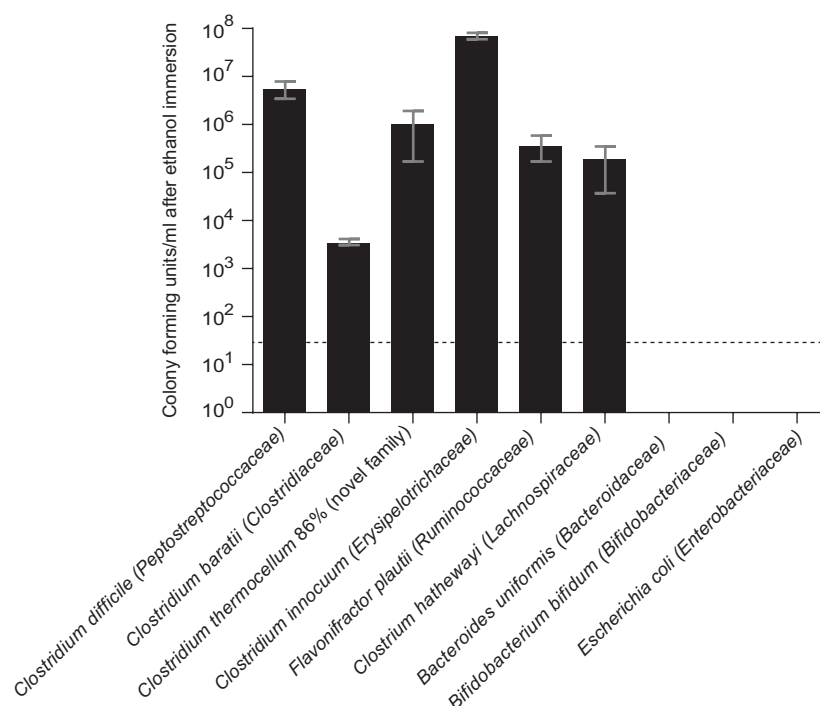
Extended Data Figure 5 | Genomic signature of sporulation within the human intestinal microbiome. A genomic signature for identifying spore-forming bacterial species contains sporulation- and germination-associated genes and genes not previously associated with sporulation. Characterized sporulation genes are on the outer circle, genes not associated with a specific sporulation cycle or uncharacterized genes are in the inside rectangle. *C. difficile* strain 630 gene names are used when possible, otherwise locus tag identifiers are shown. *Bacillus subtilis* gene

names are used when no *C. difficile* homologue is available. The signature is enriched with known sporulation-associated genes from stages I–V of the spore formation and germination cycles (significant at $q < 3.0 \times 10^{-37}$, Fisher's exact test). Genes associated with regulation are present with at least 10 genes coding for regulatory or DNA-binding roles ($q < 1.4 \times 10^{-5}$, Fisher's exact test). Genes not previously associated with sporulation are also present and these have putative roles as heat shock, membrane-associated proteins and DNA-polymerase-associated proteins.



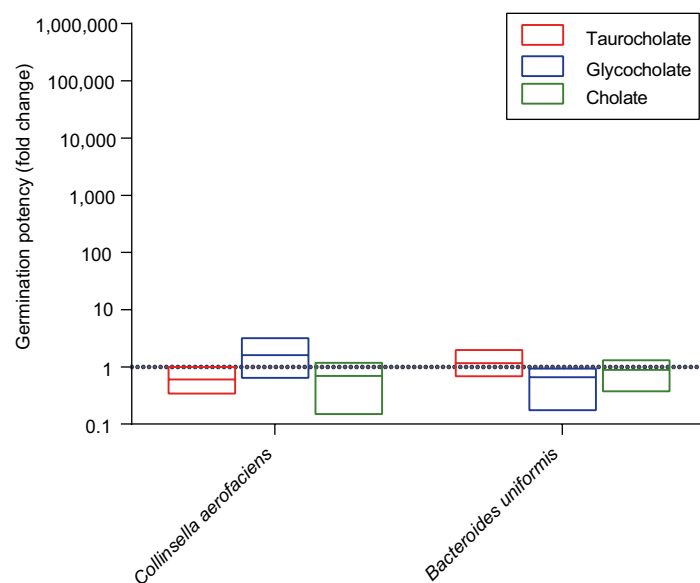
Extended Data Figure 6 | Validation and characterization of the sporulation signature. **a**, The signature accurately distinguishes spore-forming and non-spore-forming bacteria cultured from this study and from across different environments (known spore-formers $n = 57$, known non-spore-formers $n = 50$, cultured after ethanol treatment $n = 69$,

cultured after no ethanol treatment $n = 149$). Refer to Supplementary Table 1 for signature scores of the bacteria tested. Mean \pm s.d. **b**, Assignment of functional classes to the signature reveals a wide range of functional processes with sporulation- and regulation-associated genes dominating.



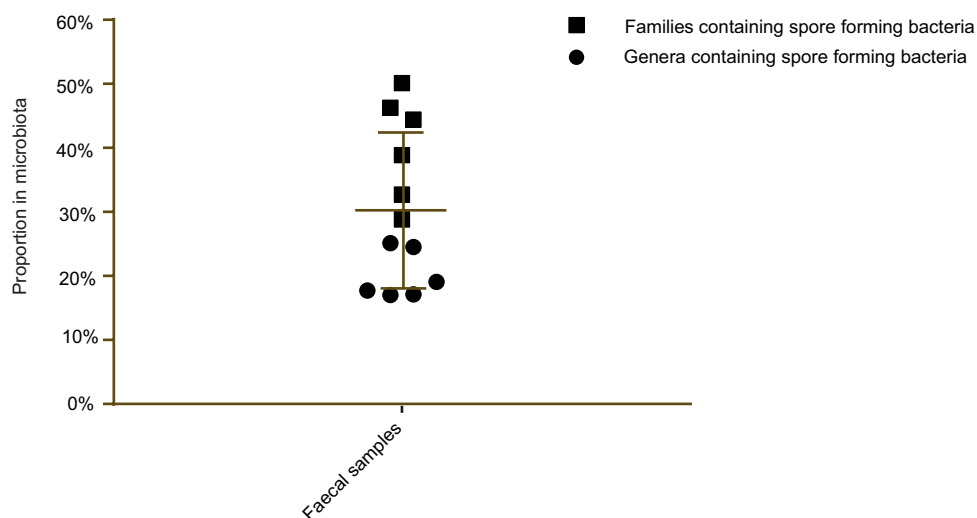
Extended Data Figure 7 | Spore-forming bacteria are more resilient than non-spore-forming bacteria to environmental stresses such as disinfectants. Pure bacterial cultures were immersed in ethanol for 4 h before being washed and inoculated onto YCFA growth medium

with sodium taurocholate as a germinant. Only spore-forming bacteria survived. Taxonomic family names are shown in brackets. The dashed line indicates the culture detection limit of 50 c.f.u. ml⁻¹. Mean ± s.d., $n = 3$ biological replicates for each species tested.



Extended Data Figure 8 | Growth response of non-spore-forming bacteria to intestinal germinants. The number of c.f.u. present on plates in the presence of a particular germinant expressed as a fold change with respect to the number of c.f.u. present on plates in the absence of a germinant. No ethanol shock treatment was performed beforehand.

A fold change of one (dashed line) would indicate that a germinant had no effect on the number of c.f.u. recovered from the bacteria. There was no statistically significant difference based on an unpaired *t*-test of each germinant condition against the no germinant condition. Mean and range, $n = 3$ biological replicates for both species.



Extended Data Figure 9 | Validation of the estimation of the proportion of spore-formers in the intestinal microbiota. Full-length 16S rRNA gene amplicon sequencing was used to determine the taxonomic proportions of bacteria from the six donor faecal samples. Spore-forming bacteria were cultured from each donor and a taxonomic classification

was assigned as described in the main text. The genus (circle) and family (square) taxonomic ranks were designated as the lower and upper limits for calculating the proportion of spore-formers at a taxonomic level. Specific genera and families were included if they contained a species that was cultured after ethanol shock treatment. Mean \pm s.d.

Redirecting abiraterone metabolism to fine-tune prostate cancer anti-androgen therapy

Zhenfei Li¹, Mohammad Alyamani¹, Jianneng Li¹, Kevin Rogacki², Mohamed Abazeed², Sunil K. Upadhyay³, Steven P. Balk⁴, Mary-Ellen Taplin⁵, Richard J. Auchus³ & Nima Sharifi^{1,6,7}

Abiraterone blocks androgen synthesis and prolongs survival in patients with castration-resistant prostate cancer, which is otherwise driven by intratumoral androgen synthesis^{1,2}. Abiraterone is metabolized in patients to Δ^4 -abiraterone (D4A), which has even greater anti-tumour activity and is structurally similar to endogenous steroidal 5 α -reductase substrates, such as testosterone³. Here, we show that D4A is converted to at least three 5 α -reduced and three 5 β -reduced metabolites in human serum. The initial 5 α -reduced metabolite, 3-keto-5 α -abiraterone, is present at higher concentrations than D4A in patients with prostate cancer taking abiraterone, and is an androgen receptor agonist, which promotes prostate cancer progression. In a clinical trial of abiraterone alone, followed by abiraterone plus dutasteride (a 5 α -reductase inhibitor), 3-keto-5 α -abiraterone and downstream metabolites were depleted by the addition of dutasteride, while D4A concentrations rose, showing that dutasteride effectively blocks production of a tumour-promoting metabolite and permits D4A accumulation. Furthermore, dutasteride did not deplete the three 5 β -reduced metabolites, which were also clinically detectable, demonstrating the specific biochemical effects of pharmacological 5 α -reductase inhibition on abiraterone metabolism. Our findings suggest a previously unappreciated and biochemically specific method of clinically fine-tuning abiraterone metabolism to optimize therapy.

Metastatic prostate cancer generally responds initially to medical or surgical castration but eventually develops into castration-resistant prostate cancer (CRPC), which is driven by the metabolic capability of tumours to reconstitute potent androgens, mainly from dehydroepiandrosterone (DHEA)/DHEA-sulfate, and in turn stimulate the androgen receptor^{1,4,5}. Abiraterone (Abi; administered orally as Abi acetate), a steroidal drug, inhibits 17 α -hydroxylase/17,20-lyase (CYP17A1), blocks androgen synthesis and prolongs survival, even after treatment with docetaxel chemotherapy^{6,7}. Unfortunately, disease progression ultimately results in tumour lethality.

Abi is converted in patients by 3 β -hydroxysteroid dehydrogenase (3 β HSD) to D4A, which blocks multiple enzymes required for 5 α -dihydrotestosterone (DHT) synthesis, directly and potently antagonizes the androgen receptor, and has more potent anti-tumour activity than abiraterone itself³. However, there is no known method to increase accumulation of D4A as an Abi metabolite and it is not known whether other Abi metabolites harbour clinically relevant biochemical activity and contribute to response or resistance to treatment with Abi.

The Δ^4 ,3-keto structure of D4A makes it potentially susceptible to 5 α -reduction to 3-keto-5 α -Abi (5 α -Abi) or 5 β -reduction to 3-keto-5 β -Abi (5 β -Abi), both of which are irreversible reactions (Fig. 1a). 3-keto-reduction of both of these metabolites can reversibly convert them to their 3 α -OH and 3 β -OH congeners, making a total of six

metabolites downstream of D4A (Fig. 1a and Extended Data Fig. 1). Mass spectrometry showed conversion from Abi and D4A to all three 5 α -reduced metabolites, interconversion among the three 5 α -reduced metabolites, and interconversion among the three 5 β -reduced metabolites in the LAPC4, C4-2 and VCaP prostate cancer cell lines (Fig. 1b and Extended Data Fig. 2). Direct incubation of LNCaP and LAPC4 human prostate cancer cells with D4A resulted in conversion to 5 α -Abi and 3 α -OH-5 α -Abi, as detected using high-performance liquid chromatography (HPLC) with ultraviolet absorption (Extended Data Fig. 3a, d), and treatment with 5 α -Abi yielded conversion to 3 α -OH-5 α -Abi (Extended Data Fig. 3b). The reversibility of this reaction is shown by detection of 5 α -Abi after 3 α -OH-5 α -Abi treatment, particularly in LAPC4 cells (Extended Data Fig. 3c); however, reduction to 3 α -OH-5 α -Abi appears to be the preferred reaction direction. Similarly, in mice, 5 α -Abi was preferentially converted to 3 α -OH-5 α -Abi, although the reverse reaction was also detected (Fig. 1c). 3 β -OH-5 α -Abi was also oxidized to 5 α -Abi and converted to 3 α -OH-5 α -Abi. Reflecting the irreversible nature of steroid 5 α -reduction, no Abi, D4A, or 5 β -reduced metabolites were detected in mice after treatment with any of the 5 α -reduced Abi metabolites. Furthermore, all six metabolites were detected in the sera of 12 patients with CRPC undergoing treatment with Abi acetate (Fig. 1d, Extended Data Fig. 4 and Extended Data Table 1). Together, these data support a model in which once D4A is 5 α -reduced to 5 α -Abi, 3-keto reduction to 3-(α and β)-OH isomers and the reverse reactions occur, both in prostate cancer cells and *in vivo*.

Steroid 5 α -reduction preserves the steroid planar structure and is essential for the regulation of biologically active androgens (that is, conversion of testosterone to DHT and Δ^4 -androstenedione to 5 α -androstenedione (5 α -dione))^{8,9}. On the other hand, steroid 5 β -reduction disrupts the planar conformation by introducing a 90° bend, which generally inactivates steroid hormones and facilitates their clearance. We therefore focused subsequent studies on the pathway and metabolites of D4A 5 α -reduction. Synthesis of 5 α -Abi and 3 α -OH-5 α -Abi is facilitated by upstream conversion of Abi to D4A by 3 β HSD (Extended Data Fig. 5a). In cells without endogenous steroid-5 α -reductase (SRD5A) expression, conversion of D4A to 5 α -Abi is enabled by expression of either SRD5A1 or SRD5A2 (Extended Data Fig. 5b). In LAPC4 cells, which predominantly express SRD5A1 (ref. 8), genetically silencing SRD5A1 (Extended Data Fig. 5c), or pharmacological blockade using the SRD5A1 inhibitor LY191704 (ref. 10) or clinically achievable concentrations of the dual isoenzyme inhibitor dutasteride¹¹, blocked conversion of D4A to 5 α -Abi and 3 α -OH-5 α -Abi (Extended Data Fig. 5d). The aldo-keto reductase isoenzyme AKR1C2 is thought to be the predominant 3-keto reductase that converts the 3-keto steroid DHT to 5 α -androstane-3 α ,17 β -diol¹². We found that AKR1C2 expression also enabled the reduction of 5 α -Abi to 3 α -OH-5 α -Abi (Extended Data Fig. 5e).

¹Department of Cancer Biology, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio 44195, USA. ²Department of Radiation Oncology, Taussig Cancer Institute, Cleveland Clinic, Cleveland, Ohio 44195, USA. ³Departments of Pharmacology and Internal Medicine, Division of Endocrinology and Metabolism, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA.

⁴Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts 02215, USA. ⁵Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts 02115, USA.

⁶Department of Urology, Glickman Urological and Kidney Institute, Cleveland Clinic, Cleveland, Ohio 44195, USA. ⁷Department of Hematology and Oncology, Taussig Cancer Institute, Cleveland Clinic, Cleveland, Ohio 44195, USA.

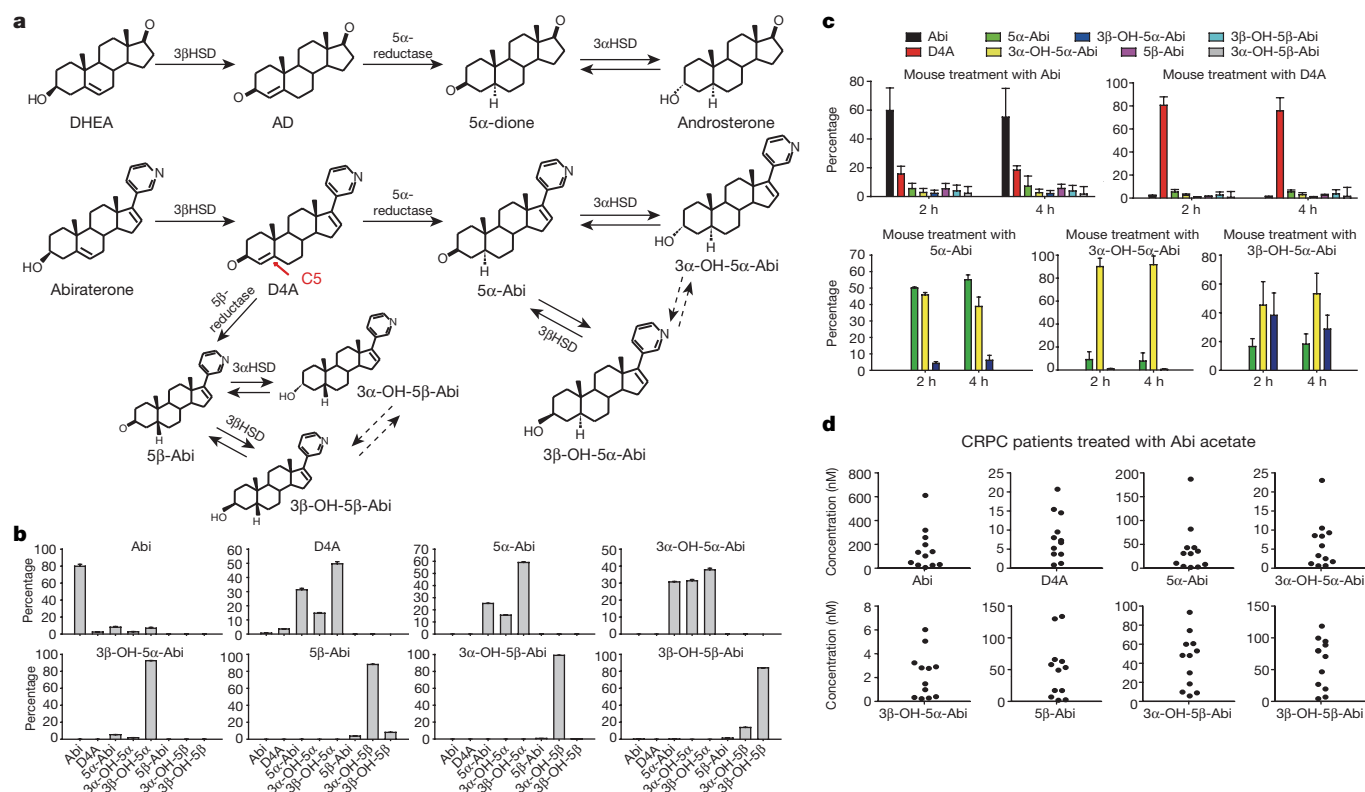


Figure 1 | Genesis of 5 α - and 5 β -reduced Abi metabolites in patients treated with Abi acetate. **a**, Abiraterone (Abi) metabolism by steroidogenic enzymes. Top, in a pathway of androgen metabolism, DHEA is converted by 3 β HSD to Δ^4 -androstenedione (AD), which is 5 α -reduced to 5 α -androstenedione (5 α -dione), which is in turn 3-keto-reduced to androsterone. Bottom, structurally analogous conversion of abiraterone to D4A enables 5 α - and 5 β -reduction of D4A at C5 (red arrow), yielding six additional Abi metabolites. Dotted arrows indicate uncertainty of direct conversion between 3 α - and 3 β -isomers (compared to indirect conversion via the 3-keto intermediate). **b**, Interconversion of Abi metabolites in the LAPC4 prostate cancer cell line. Cells were treated

with Abi or the indicated metabolite (0.1 μ M) for 48 h and each of the indicated metabolites was detected by liquid chromatography–tandem mass spectrometry (LC–MS/MS) in triplicate. **c**, *In vivo* Abi metabolism in mice. Treatment with Abi ($n = 5$ mice) or D4A ($n = 5$ mice) results in detection of all six 5-reduced metabolites. Treatment with any of the three 5 α -reduced Abi metabolites ($n = 4$ mice for each compound) results in detection of the other two 5 α -reduced metabolites, demonstrating interconversion. Error bars represent s.d. in **b** and **c**. **d**, Abi metabolites in sera of 12 patients with prostate cancer treated with Abi acetate. Metabolites were measured by LC–MS/MS.

Next, we investigated the effects of 5 α -reduced Abi metabolites on the androgen pathway. 5 α -reduction of D4A to 5 α -Abi and 3 α -OH-5 α -Abi is accompanied by attenuation or loss of inhibition of CYP17A1, 3 β HSD and SRD5A (Fig. 2a–c), as assessed by conversion of [3 H]pregnenolone to DHEA, [3 H]DHEA to Δ^4 -androstenedione, and [3 H] Δ^4 -androstenedione to 5 α -dione, respectively. The effects, or lack thereof, of D4A and 5 α -Abi metabolites on 3 β HSD are consistent with previous observations that endogenous Δ^4 , 3-keto-steroids inhibit 3 β HSD and that 5 α -reduction of Δ^4 , 3-keto-steroids leads to loss of their inhibitory activity¹³.

D4A and 5 α -Abi have similar affinities for the Thr877Ala mutant androgen receptor in LNCaP cells and the wild-type androgen receptor in LAPC4 cells, whereas 3 α -OH-5 α -Abi and Abi have lower affinities (Fig. 2d). To assess the consequences of 5 α -Abi binding to the androgen receptor, we measured expression of androgen-responsive genes. Treatment with 5 α -Abi resulted in expression of androgen-responsive genes in LAPC4 and LNCaP cells (Fig. 2e–g). These genes were expressed to a lower level after 3 α -OH-5 α -Abi treatment. The delayed and modest effect of 3 α -OH-5 α -Abi on induction of prostate-specific antigen (PSA) is consistent with its low binding affinity for the androgen receptor and the modest conversion of 3 α -OH-5 α -Abi to 5 α -Abi, that appears to occur to a greater extent in LAPC4 than in LNCaP cells (Extended Data Fig. 3c). Using a cDNA microarray and Gene Set Enrichment Analysis we found that 5 α -Abi stimulates androgen receptor signature genes¹⁴ (Extended Data Fig. 6a, b) and that 81% of genes regulated by 5 α -Abi are also regulated by

DHT (Extended Data Fig. 6c, d and Extended Data Table 2). To test the effect of androgen receptor stimulation by 5 α -Abi on tumour growth, we treated mice bearing CRPC xenografts with 5 α -Abi or 3 α -OH-5 α -Abi. 5 α -Abi significantly shortened progression-free survival ($P < 0.01$), whereas 3 α -OH-5 α -Abi had no detectable effect when compared to untreated mice (Fig. 2h). We also tested all three 5 β -reduced Abi metabolites for effects on androgen-responsive gene expression and confirmed that perturbation of the steroid planar structure is accompanied by the absence of metabolite activity (Fig. 2i).

We hypothesized that, when treated with Abi, prostate cancer cells might develop the capacity to augment the conversion of D4A to 5 α -Abi by upregulating SRD5A. VCaP and LNCaP cells were cultured for 6 months with D4A or Abi, along with DHEA, to mimic the human adrenal androgen milieu (Fig. 3a). Cells propagated in long-term culture with D4A and Abi exhibited increased SRD5A enzyme activity, as assessed by conversion of [3 H] Δ^4 -androstenedione to 5 α -reduced androgens, [3 H]testosterone to DHT, and D4A to 5 α -reduced Abi metabolites (Fig. 3b, c). Increased SRD5A enzyme activity was accompanied by a predominant increase in SRD5A1 mRNA (Extended Data Fig. 7a) and protein expression (Fig. 3d, e). Unlike Abi or D4A treatment, enzalutamide treatment did not induce an increase in SRD5A (Extended Data Fig. 7b, c). No change in AR-V7 (the androgen receptor splice variant) resulted from Abi or D4A treatment (Extended Data Fig. 7d). Together, these results suggest that a selective growth advantage results from depleting the anti-tumour

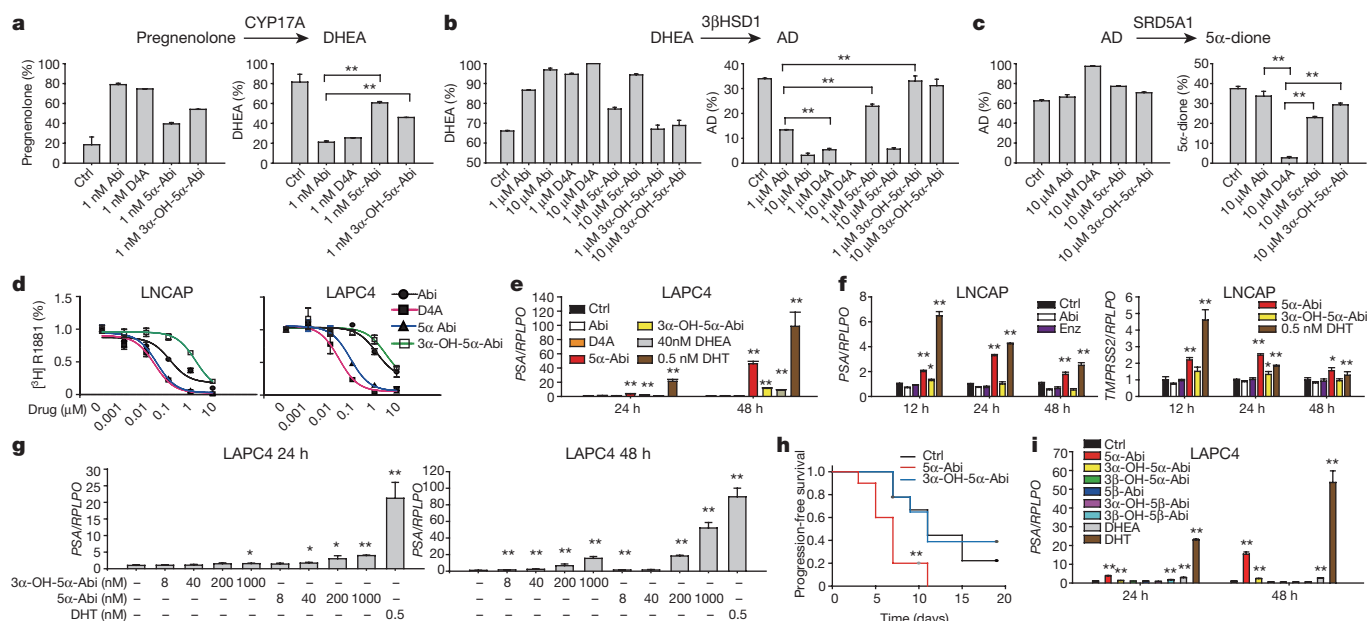


Figure 2 | Effects of 5 α -reduced Abi metabolites on the androgen pathway and tumour progression. **a**, Potent inhibition of CYP17A1 by Abi and D4A is attenuated upon conversion to 5 α -Abi and 3 α -OH-5 α -Abi. 293 cells overexpressing CYP17A1 were treated with [3 H]pregnenolone and conversion to DHEA was assessed in the presence of the indicated drugs. **b**, 5 α -reduction of D4A causes loss of inhibition of 3 β HSD. LNCaP cells were treated with [3 H]DHEA and the indicated drugs for 48 h, and metabolic flux to AD was assessed. **c**, D4A-mediated inhibition of SRD5A is lost upon conversion to 5 α -Abi and 3 α -OH-5 α -Abi. LAPC4 cells were treated with [3 H]AD and the indicated drugs, and flux to 5 α -dione was assessed after 24 h of incubation. **d**, 5 α -Abi and D4A bind comparably to the androgen receptor. LNCaP and LAPC4 express the mutated and wild-type androgen receptor, respectively, and were incubated with [3 H]R1881 and the indicated compounds for 30 min. Intracellular radioactivity was normalized to

protein concentration. **e**, **f**, 5 α -Abi stimulates androgen-responsive gene expression in LAPC4 and LNCaP cells. Delayed and more modest PSA (also known as *KLK3*) expression is stimulated by 3 α -OH-5 α -Abi. **g**, Dose-dependent stimulation of PSA expression by 5 α -Abi in LAPC4 cells. **h**, Treatment with 5 α -Abi ($n = 10$ mice) but not 3 α -OH-5 α -Abi ($n = 9$ mice) hastens VCaP CRPC xenograft growth compared with control treatment ($n = 9$ mice) in orchiectomized mice. Treatment with the indicated compounds began when CRPC tumours reached 100 mm 3 and progression-free survival was assessed as the time until there was >30% growth for two sequential measurements. **i**, 5 β -reduced Abi metabolites do not stimulate PSA expression. Expression is normalized to RPLP0 and vehicle expression in **e–g** and **i**. All error bars represent s.d. All experiments in **a–g** and **i** were performed at least three times.

activity associated with D4A and/or increasing the androgen receptor agonist activity of 5 α -Abi.

Next, we hypothesized that the increased 5 α -Abi:D4A ratio (approximately 2.5:1) would be specifically and clinically reversible by dual SRD5A isoenzyme inhibition with dutasteride in patients being treated with Abi acetate. A phase II clinical trial (NCT01393730) involving treatment with Abi acetate (1,000 mg daily) plus prednisone (5 mg daily) for 2 months (2 cycles), followed by the addition of dutasteride at the start of cycle 3 (3.5 mg once daily; Fig. 4a), is ongoing in men with metastatic CRPC. The results from sixteen patients who had blood collected on Abi acetate alone (start of cycle 3) and after the addition of dutasteride (start of cycles 4 and 7) have been analysed. Notably, there was an 89% decline in the mean concentration of 5 α -Abi after the addition of dutasteride (cycle 3: 25.8 nM versus cycle 4: 2.9 nM; $P < 0.001$; Fig. 4d). The other two 5 α -reduced metabolites downstream of SRD5A exhibited similar declines (92% decline in 3 α -OH-5 α -Abi and 73% decline in 3 β -OH-5 α -Abi), further corroborating the ability of dutasteride to block 5 α -reduction of D4A in patients. Pharmacological inhibition of SRD5A nearly doubled the mean serum concentration of D4A (cycle 3: 9.9 nM versus cycle 4: 18.2 nM; $P = 0.002$; Fig. 4c). Unexpectedly, the addition of dutasteride also nearly doubled the mean concentration of Abi (cycle 3: 191.2 nM versus cycle 4: 372.4 nM; $P = 0.051$; Fig. 4b), although this difference did not reach statistical significance. Concentrations of Abi, D4A and 5 α -Abi metabolites at cycle 7, the second time point after addition of dutasteride, were similar to those at cycle 4, although the differences from Abi alone at baseline were slightly lessened (Fig. 4 and Extended Data Table 3). Finally, and in sharp contrast to the substantial decline in 5 α -Abi metabolites after the addition of dutasteride, there was no

decrease in any of the three 5 β -reduced Abi metabolites, supporting the idea that SRD5A inhibition has a very specific biochemical effect on 5 α -Abi metabolism (Fig. 4e). Together, these findings demonstrate that the elevated 5 α -Abi:D4A ratio can be pharmacologically, specifically and clinically reversed by dutasteride.

CYP17A1 inhibition by Abi is clinically incomplete, as has been demonstrated by the persistence of residual urinary androgen metabolites¹⁵ and high residual serum concentrations of DHEA-sulfate, the major androgen produced from the human adrenal gland¹⁶, in patients being treated with Abi. Molecular aberrations that sustain androgen receptor signalling contribute to the development of CRPC¹⁷ and also drive Abi resistance^{18,19}. Together, these studies suggest that reversal of sustained androgen receptor signalling should have a therapeutic benefit in at least a subset of patients with Abi resistant disease.

Some sustained androgen receptor signalling is due to the continuous supply of endogenous androgens (testosterone and/or DHT) provided by maintained steroidogenesis. Preclinical models suggest that Abi resistance can be driven by upregulation of steroidogenic enzymes²⁰. Our findings demonstrate that, as well as providing potent endogenous androgens, steroidogenic enzymes may also regulate Abi metabolism, for example by increasing SRD5A activity and thereby hastening the elimination of the anti-tumour activity associated with D4A (ref. 3) by converting it to 5 α -Abi, which instead has tumour-promoting androgen receptor agonist activity. Consistent with our findings, SRD5A is one of the most upregulated steroidogenic-enzyme transcripts in another model of Abi resistance²⁰.

The coordinate effects of steroidogenic enzymes on endogenous steroids, compared to the effects on Abi, should be considered in view

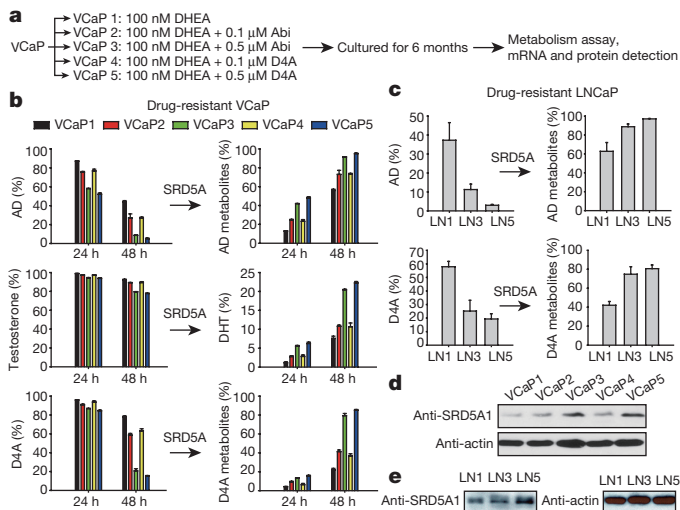


Figure 3 | Long-term exposure to Abi and D4A leads to an increase in SRD5A expression and enzymatic activity and an increase in conversion from D4A to 5 α -reduced Abi metabolites. **a**, Experimental schema. VCaP cells were cultured with DHEA (100 nM) alone or with the indicated concentration of Abi or D4A continuously for 6 months. LNCaP cells were cultured under similar conditions (LN1, LN3 and LN5 correspond to VCaP1, VCaP3 and VCaP5 treatment, respectively). **b**, **c**, Treatment with Abi or D4A induces an increase in SRD5A enzyme activity and increased conversion of D4A to 5 α -Abi metabolites in VCaP (**b**) and LNCaP (**c**) cells. Cells were treated with [3 H]androstenedione (AD), [3 H]testosterone (T), or D4A for 24 or 48 h, and conversion to 5 α -reduced metabolites was assessed by HPLC. **d**, **e**, SRD5A1 protein expression is increased by Abi and D4A treatment in VCaP (**d**) and LNCaP (**e**) cells. Experiments in **b** and **c** were performed in triplicate; error bars represent s.d. All experiments were performed at least three times.

of our findings on extensive steroidogenic metabolism of Abi. For example, increased SRD5A enzyme activity should have concordant favourable effects on tumour growth by increasing DHT synthesis and converting D4A to 5 α -Abi. On the other hand, increasing 3 β HSD activity would be expected to have discordant effects, as it is required for synthesis of testosterone and DHT (which promote tumour growth) but increases the conversion of Abi to D4A (which is detrimental to tumour growth). In this context, it is likely that the net effect of increased intratumoural 3 β HSD activity will be beneficial, because the endogenous substrates (DHEA, Δ^5 -androstenediol and pregnenolone) are probably preferred over D4A. A third enzymatic reaction that is relevant both to endogenous 5 α -reduced androgens and 5 α -Abi is 3 α -OH-oxidation to 3-keto-steroids. Oxidation or 'back-conversion' of 3 α -OH-5 α -androstenediol, which does not stimulate the androgen receptor, to 3-keto-DHT can stimulate androgen receptor signalling^{21,22}. The net effect of this reaction on androgens and Abi metabolism is therefore expected to be concordant and stimulatory because, in addition to increasing DHT synthesis, conversion of 3 α -OH-5 α -Abi to 5 α -Abi also increases affinity for the androgen receptor (Fig. 2d), and stimulates androgen-responsive gene expression (Fig. 2e–g) and tumour growth (Fig. 2h).

Notwithstanding these considerations, the majority of Abi metabolites found in serum probably form independently of tumour metabolism and result from the activity of hepatic enzymes. Although we found that prostate cancer cell lines readily 5 α -reduce D4A, no prostate cancer-dependent 5 β -reduction was observed. Both steroid 5 α - and 5 β -reductase reactions are abundant in the liver. The contributions to treatment response of D4A depletion by intrinsic tumour SRD5A versus extrinsic metabolism remain to be determined. Nonetheless, in our clinical study, adding dutasteride to treatment with Abi acetate resulted in an approximately 90% decline in circulating concentrations of 5 α -Abi, the immediate product of D4A 5 α -reduction, with

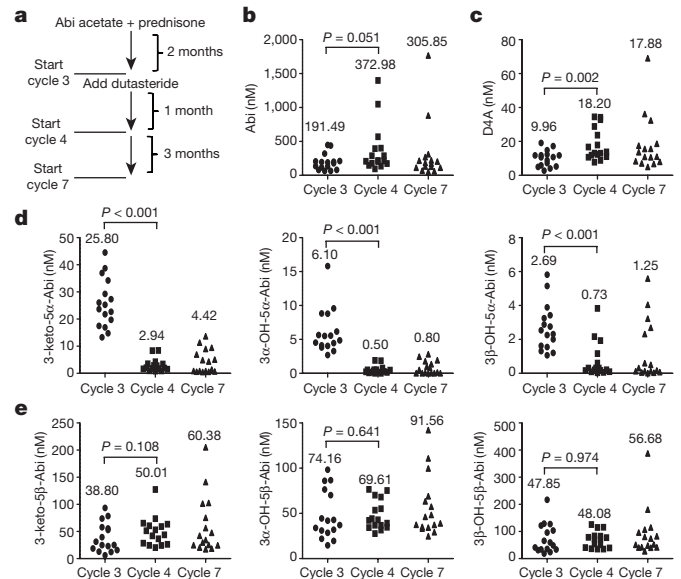


Figure 4 | In patients treated with Abi acetate, SRD5A inhibition significantly increases serum D4A and specifically and significantly depletes all three 5 α -Abi metabolites in serum. **a**, Clinical trial schema. Blood was tested for Abi metabolites after 2 months of treatment with Abi acetate and prednisone (start cycle 3), and again after 1 and 4 months (start cycles 4 and 7) of addition of treatment with dutasteride. **b**–**e**, Serum concentrations of Abi (**b**), D4A (**c**), all three 5 α -reduced Abi metabolites (**d**) and all three 5 β -reduced Abi metabolites (**e**). Numbers indicate mean Abi and Abi metabolite concentrations at each of the three time points.

similar declines in the other 5 α -Abi metabolites. D4A concentrations concomitantly rose after the addition of dutasteride, supporting the specific pharmacological effect of dutasteride on D4A metabolism. Numerically, there was also a rise in Abi concentration, and although this increase did not reach the generally accepted level of statistical significance ($P = 0.051$), it does raise the possibility that SRD5A provides an important mechanism of Abi clearance, at least in a subset of patients.

The absence of any effect on 5 β -reduced Abi metabolites demonstrates the remarkable specificity of dutasteride on 5 α -reduction of D4A. The presence and maintenance of 5 β -reduced metabolites further suggests an 'escape' mechanism of D4A metabolism that occurs with pharmacological 5 α -reductase inhibition, raising the possibility that concentrations of D4A and perhaps Abi might be further elevated by 5 β -reductase inhibition, resulting in further therapeutic effects.

Our studies demonstrate that SRD5A inhibition has a clear and specific biochemical effect on D4A metabolism in patients treated with Abi. This effect would be expected to intensify the benefit of Abi therapy. The clinical benefit of fine-tuning Abi metabolism with SRD5A inhibition requires further investigation in randomized trials.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 26 September 2015; accepted 23 March 2016.

- Attard, G. *et al.* Prostate cancer. *Lancet* **387**, 70–82 (2016).
- Sharifi, N. Mechanisms of androgen receptor activation in castration-resistant prostate cancer. *Endocrinology* **154**, 4010–4017 (2013).
- Li, Z. *et al.* Conversion of abiraterone to D4A drives anti-tumour activity in prostate cancer. *Nature* **523**, 347–351 (2015).
- Chang, K. H. *et al.* A gain-of-function mutation in DHT synthesis in castration-resistant prostate cancer. *Cell* **154**, 1074–1084 (2013).
- Scher, H. I. & Sawyers, C. L. Biology of progressive, castration-resistant prostate cancer: directed therapies targeting the androgen-receptor signaling axis. *J. Clin. Oncol.* **23**, 8253–8261 (2005).

6. de Bono, J. S. *et al.* Abiraterone and increased survival in metastatic prostate cancer. *N. Engl. J. Med.* **364**, 1995–2005 (2011).
7. Ryan, C. J. *et al.* Abiraterone in metastatic prostate cancer without previous chemotherapy. *N. Engl. J. Med.* **368**, 138–148 (2013).
8. Chang, K. H. *et al.* Dihydrotestosterone synthesis bypasses testosterone to drive castration-resistant prostate cancer. *Proc. Natl Acad. Sci. USA* **108**, 13728–13733 (2011).
9. Russell, D. W. & Wilson, J. D. Steroid 5 α -reductase: two genes/two enzymes. *Annu. Rev. Biochem.* **63**, 25–61 (1994).
10. Hirsch, K. S. *et al.* LY191704: a selective, nonsteroidal inhibitor of human steroid 5 α -reductase type 1. *Proc. Natl Acad. Sci. USA* **90**, 5277–5281 (1993).
11. Clark, R. V. *et al.* Marked suppression of dihydrotestosterone in men with benign prostatic hyperplasia by dutasteride, a dual 5 α -reductase inhibitor. *J. Clin. Endocrinol. Metab.* **89**, 2179–2184 (2004).
12. Rižner, T. L. *et al.* Human type 3 3 α -hydroxysteroid dehydrogenase (aldo-keto reductase 1C2) and androgen metabolism in prostate cells. *Endocrinology* **144**, 2922–2932 (2003).
13. Byrne, G. C., Perry, Y. S. & Winter, J. S. Steroid inhibitory effects upon human adrenal 3 β -hydroxysteroid dehydrogenase activity. *J. Clin. Endocrinol. Metab.* **62**, 413–418 (1986).
14. Arora, V. K. *et al.* Glucocorticoid receptor confers resistance to antiandrogens by bypassing androgen receptor blockade. *Cell* **155**, 1309–1322 (2013).
15. Attard, G. *et al.* Clinical and biochemical consequences of CYP17A1 inhibition with abiraterone given with and without exogenous glucocorticoids in castrate men with advanced prostate cancer. *J. Clin. Endocrinol. Metab.* **97**, 507–516 (2012).
16. Taplin, M. E. *et al.* Intense androgen-deprivation therapy with abiraterone acetate plus leuprolide acetate in patients with localized high-risk prostate cancer: results of a randomized phase II neoadjuvant study. *J. Clin. Oncol.* **32**, 3705–3715 (2014).
17. Robinson, D. *et al.* Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215–1228 (2015).
18. Carreira, S. *et al.* Tumor clone dynamics in lethal prostate cancer. *Sci. Transl. Med.* **6**, 254ra125 (2014).
19. Miyamoto, D. T. *et al.* Androgen receptor signaling in circulating tumor cells as a marker of hormonally responsive prostate cancer. *Cancer Discov.* **2**, 995–1003 (2012).
20. Mostaghel, E. A. *et al.* Resistance to CYP17A1 inhibition with abiraterone in castration-resistant prostate cancer: induction of steroidogenesis and androgen receptor splice variants. *Clin. Cancer Res.* **17**, 5913–5925 (2011).
21. Biswas, M. G. & Russell, D. W. Expression cloning and characterization of oxidative 17 β - and 3 α -hydroxysteroid dehydrogenases from rat and human prostate. *J. Biol. Chem.* **272**, 15959–15966 (1997).
22. Mohler, J. L. *et al.* Activation of the androgen receptor by intratumoral bioconversion of androstenediol to dihydrotestosterone in prostate cancer. *Cancer Res.* **71**, 1486–1496 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank T. Penning for use of the AKR1C2 construct and D. Russell for LY191704. This work was supported in part by funding from a Howard Hughes Medical Institute Physician-Scientist Early Career Award (to N.S.), a Prostate Cancer Foundation Challenge Award (to N.S.), an American Cancer Society Research Scholar Award (12-038-01-CCE; to N.S.), grants from the National Cancer Institute (R01CA168899, R01CA172382, and R01CA190289; to N.S.), a grant from the US Army Medical Research and Materiel Command (PC121382 to Z.L.), a Prostate Cancer Foundation Young Investigator Award (to Z.L.), grants from the National Cancer Institute (P01 CA163227 and P50 CA090381), and a Prostate Cancer Foundation Challenge Award (to S.P.B.). Janssen provided clinical trial support (to M.-E.T.).

Author Contributions Z.L. performed gene expression, metabolism and mouse work. M. Alyamani performed mass spectrometry metabolism work. J.L. performed immunoblots. S.K.U. performed chemical syntheses. K.R. and M. Abazeed performed the microarray GSEA analysis. M.-E.T. and S.P.B. designed and performed the clinical trial. Z.L., M. Alaymani, R.J.A. and N.S. designed the studies and wrote the manuscript. All authors discussed the results and commented on the manuscript.

Author Information Microarray results have been deposited in the NCBI Gene Expression Omnibus database under accession number GSE75387. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to N.S. (sharifn@ccf.org).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Cell lines, drugs and constructs. LNCaP, 293T and VCaP cells were purchased from the American Type Culture Collection and maintained in RPMI-1640 (LNCaP) or DMEM (293T and VCaP) with 10% FBS (Gemini bio-products). The LAPC4 cell line was provided by C. Sawyers and grown in Iscove's modified Dulbecco's medium with 10% FBS. Stable LAPC4 cell lines with SRD5A1 knockdown were established as previously described⁸. Plasmid pcDNA3-c17 (a gift from W. Miller) was used to establish the 293T stable cell line expressing human *CYP17A1* as described²³. Cell lines were authenticated by DDC Medical and determined to be mycoplasma free with primers 5'-ACACCATGGGAGCTGGTAAT-3' and 5'-GTTTCATCGACTTTCAGACCCAAGGCAT-3'. Dutasteride was purchased from Medkoo Biosciences. Enzalutamide was obtained from Medivation. The AKR1C2 expression plasmid was a gift from T. Penning. Enzalutamide-resistant cells were cultured with DMSO or 1 μ M or 10 μ M enzalutamide for more than 6 months (LAPC4) or 10 weeks (VCaP) in the presence of 10 nM DHEA. Abi or D4A resistant cells were cultured as indicated in Fig. 3a.

High-performance liquid chromatography. To measure cell line metabolism, 0.2 million cells per ml were seeded and incubated in 12-well plates for ~24 h before incubation with the indicated drugs or a mixture of ³H-labelled (~1,000,000 c.p.m. per well; PerkinElmer) and non-radioactive androgens (final concentration, 100 nM) at 37 °C. Collected medium was treated with β -glucuronidase (*Helix pomatia*; Sigma-Aldrich), extracted with ethyl acetate:isooctane (1:1), and concentrated under nitrogen gas as described previously²⁴.

To measure xenograft metabolism, 10⁷ VCaP cells with Matrigel were injected subcutaneously into orchiectomized non-obese diabetic (NOD)-severe combined immunodeficient (SCID)-*Il2rg*^{-/-} (NSG) mice with 5 mg, 90-day sustained-release DHEA pellets (Innovative Research of American). Xenografts (~1,000 mm³) were removed, minced, and cultured in DMEM with 10% FBS at 37 °C with the indicated drugs. Aliquots of medium were collected at the indicated times. Collected medium was processed for HPLC with the same protocol as medium from cell lines.

HPLC analysis was performed on a Waters 717 Plus HPLC or an Agilent 1260 HPLC. Dried samples were reconstituted in 50% methanol and separated on a Kinetex 100 \times 2.1 mm, 2.6 μ m particle size C₁₈ reverse-phase column (Phenomenex) using a methanol/water gradient at 30 °C. The column effluent was analysed using a 254-nm UV-visible detector or β -RAM model 3 in-line radioactivity detector (IN/US Systems, Inc.) using Liquescent scintillation cocktail (National Diagnostics). All HPLC studies were conducted in triplicate and repeated at least three times in independent experiments. Results are shown as mean \pm s.d.

Gene expression and immunoblotting. Cells were starved with phenol red-free and serum-free medium for at least 48 h before treatment with the indicated drugs and/or androgens. RNA extraction and cDNA synthesis were performed with the GenElute Mammalian Total RNA miniprep kit (Sigma-Aldrich) and iScript cDNA Synthesis Kit (Bio-Rad), respectively. Quantitative PCR (qPCR) analysis was conducted in triplicate in an ABI 7500 Real-Time PCR machine (Applied Biosystems) using iTaq Fast SYBR Green Supermix with ROX (Bio-Rad) and primers for *TMPRSS2*, *PSA* and *RPLP0*, as described previously⁴. SYBR Premix Ex TaqII (Takara) was used for *SRD5A1*, *SRD5A2* and *ARV7* detection. Primers used for androgen receptor detection were 5'-TCTTGTCTGCTCTCGGAAATGT-3' and 5'-AAGCCTCTCCTTCCTCTCTGA-3' (ref. 25). Primers used for *ARV7* detection were 5'-CCATCTTGTCTGCTCTCGGAAATGTTATGAAGC-3' and 5'-TTTGAATGAGGCAAGTCAGCCTTCT-3' (ref. 26). Accurate quantification of each mRNA was achieved by normalizing the sample values to *RPLP0* and to vehicle-treated cells. Cell lysate (50 μ g) was used for immunoblotting with rabbit anti-SRD5A1 (Abnova) and mouse anti- β -actin (Sigma-Aldrich) antibodies.

Microarray study and analysis. LAPC4 cells were starved with phenol red-free and serum-free medium for at least 48 h before treatment with the indicated drugs and/or androgens, in biological triplicate. RNA was extracted with *mirVana* miRNA isolation kit (Life Technologies). The genomics core of Cleveland Clinic generated cDNA and performed the microarray with HumanHT-12 v4 Expression BeadChip and iScan (Illumina). Hybrid signals were analysed with Illumina GenomeStudio Software 2011.1 and normalized to the vehicle control group. Regulated genes are defined as detection $P < 0.01$, fold change (compared to control group) > 1.55 or < 0.5 . Heatmap was generated with HemI software²⁷. The complete results have been deposited in the NCBI Gene Expression Omnibus database under accession GSE75387.

Normalized, log₂-transformed data were used for subsequent gene set enrichment carried out using R (<http://www.R-project.org>) and Bioconductor software²⁸.

The 2,500 genes with the highest median absolute deviation were selected for analysis. Enrichment scores were calculated for the gene set C2 Collection (curated pathways) from the Molecular Signatures Database version 5.1 (MSigDB v5.1) using the Bioconductor package 'GSVA'²⁹. Gene sets with a minimum of 10 genes and a maximum of 1,000 genes were included. Significance testing of enrichment scores was performed with a moderated *t*-statistic (false discovery rate (FDR) < 0.01) using the Bioconductor package 'limma'. Separately, GSEA was used to correlate the 5 α -Abi expression data with an androgen receptor-selective gene set described elsewhere¹⁴. The GSEA enrichment plot was generated as described elsewhere^{30,31}.

Mouse xenograft studies. Male NSG mice, 6 to 8 weeks of age were obtained from the Cleveland Clinic Biological Resources Unit facility. All mouse studies were conducted under a protocol approved by the Cleveland Clinic Institutional Animal Care and Use Committee. 10⁷ VCaP cells were injected subcutaneously with Matrigel. Once tumours reached 100 mm³ (length \times width \times height \times 0.52), mice were surgically orchiectomized and arbitrarily assigned to vehicle ($n = 9$), 5 α -Abi ($n = 10$), or 3 α -OH-5 α -Abi ($n = 9$) treatment groups. Mice were injected intraperitoneally with 0.15 mL 5 α -Abi and 3 α -OH-5 α -Abi (0.15 mmol kg⁻¹ per day and 0.075 mmol kg⁻¹ per day, respectively, in 5% benzyl alcohol and 95% safflower oil solution) every day for up to 20 days. Control groups were administered 0.15 mL 5% benzyl alcohol and 95% safflower oil solution via intraperitoneal injection every day. Tumour volume was measured every other day, and time to increase in tumour volume by 30% was determined (2 sequential increases). Mice were killed at treatment day 20. The significance of the difference between treatment groups was assessed by Kaplan-Meier survival analysis using a log-rank test in SigmaStat 3.5.

AR competition assay. Cells were starved with phenol red-free and serum free-medium for at least 48 h and then treated with [³H]R1881 and the indicated concentrations of drugs for 30 min. Cells were washed thoroughly with PBS and then lysed with RIPA buffer. Intracellular radioactivity was measured with a Beckman Coulter LS6000IC liquid scintillation counter and normalized to the protein concentration as detected with a Wallac Victor2 1420 Multilabel counter (Perkin Elmer).

Patient serum collection and drug extraction. Twelve patients with CRPC undergoing standard treatment with Abi acetate at Cleveland Clinic were consented under an Institutional Review Board-approved protocol (Case 7813). Blood was collected using Vacutainer Plus serum blood collection tubes (#BD367814, Becton Dickinson, Franklin Lakes, NJ), between 1 and 16 h after the 1000 mg daily dose of Abi acetate was administered, and allowed to clot. Tubes were centrifuged at 2500 r.p.m. or 1430g for 10 min. Serum aliquots were frozen at -80 °C until processing. To study the effect of dutasteride on Abi metabolism, serum samples were collected from patients treated on a phase II clinical trial at Dana-Farber Cancer Institute (NCT01393730). Each patient received Abi acetate (1,000 mg daily) plus prednisone (5 mg daily) for 2 cycles (8 weeks) and then additional treatment with dutasteride (3.5 mg daily) was initiated. Samples were collected on treatment with Abi alone and after addition of dutasteride (start of cycles 3, 4 and 7). Seventeen patients treated at this institution on this clinical trial had blood available from all 3 time points. One patient had Abi concentrations that were under the limit of detection and was later determine to have stopped treatment because of adverse effects and was therefore not included in the analysis. Drug metabolites and internal standard (d₄-abiraterone, Toronto Research Chemicals Inc) were extracted from 100 μ L of patient serum with methyl tert-butyl ether (Sigma Aldrich), evaporated under a stream of nitrogen gas and reconstituted in methanol:water (50:50) before mass spectrometry analysis. Standard curves were generated using human serum spiked with known concentrations of each metabolite to enable determination of unknown concentrations in patient samples.

Mouse serum extraction. Mouse serum (20 μ L) was precipitated with 500 μ L methanol containing the internal standard (d₄-abiraterone) the supernatant was then injected into the mass analyser. Standard curves were prepared with mice serum spiked with known metabolites concentrations for accurate determination of unknown metabolites concentrations.

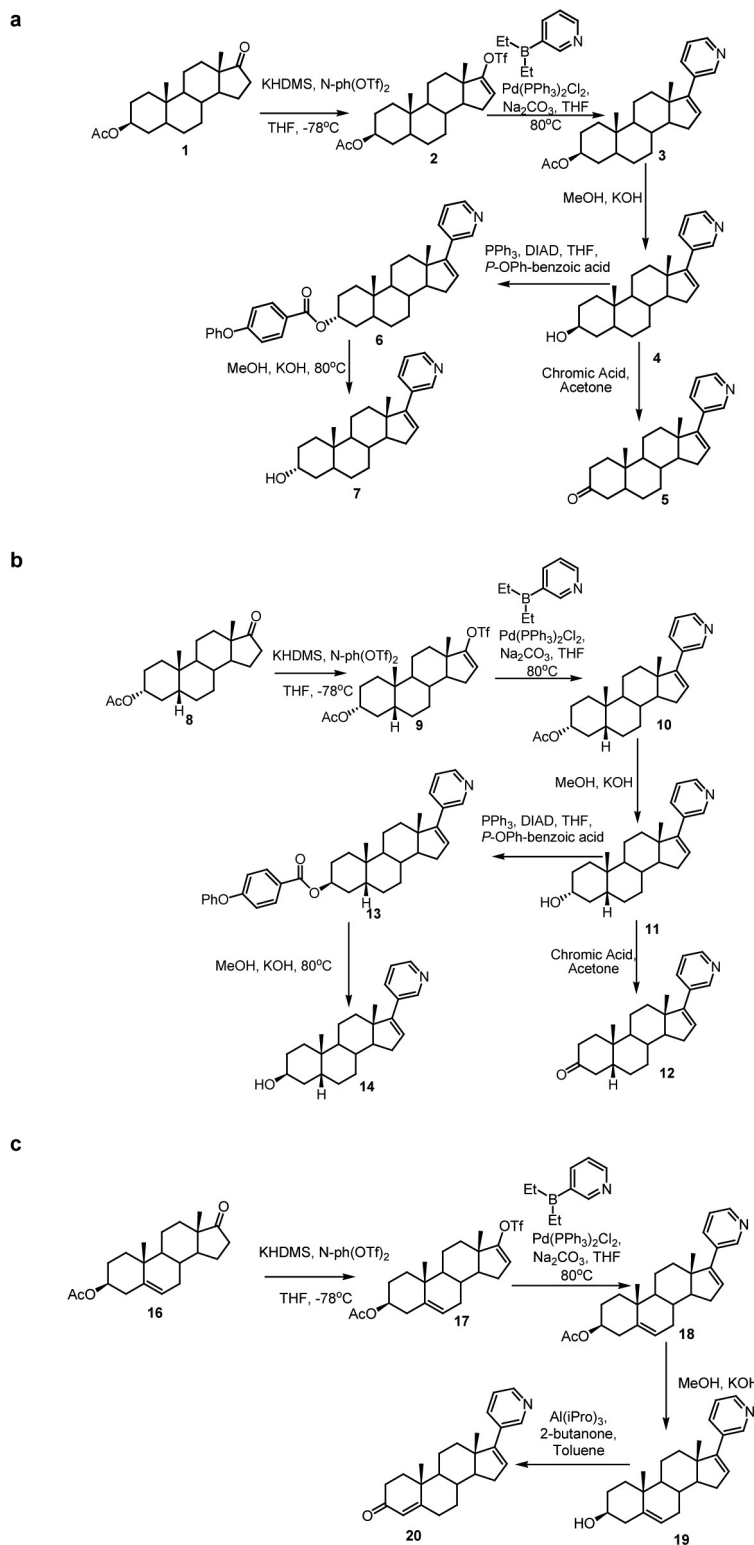
Cell line media extraction. Samples of medium (200 μ L each) collected at different time points were extracted with methyl tert-butyl ether (Sigma Aldrich), evaporated under a stream of nitrogen gas and reconstituted in methanol:water (50:50) before mass spectrometry analysis.

Mass spectrometry. Samples were analysed on a ultra high-performance liquid chromatography station (Shimadzu) with a DGU-20A3R degasser, 2 LC-30AD pumps, a SIL-30AC autosampler, a QTO-10A column oven and a CBM-20A system controller in tandem with a QTRAP 5500 mass spectrometer (AB Sciex). Drug metabolites were ionized using electrospray ionization in positive ion mode. Multiple reaction monitoring was used to follow mass transitions for Abi, IS, and the metabolites (Supplementary Table 1). Owing to the similarity in structure and mass transitions for the metabolites it was necessary to separate them with

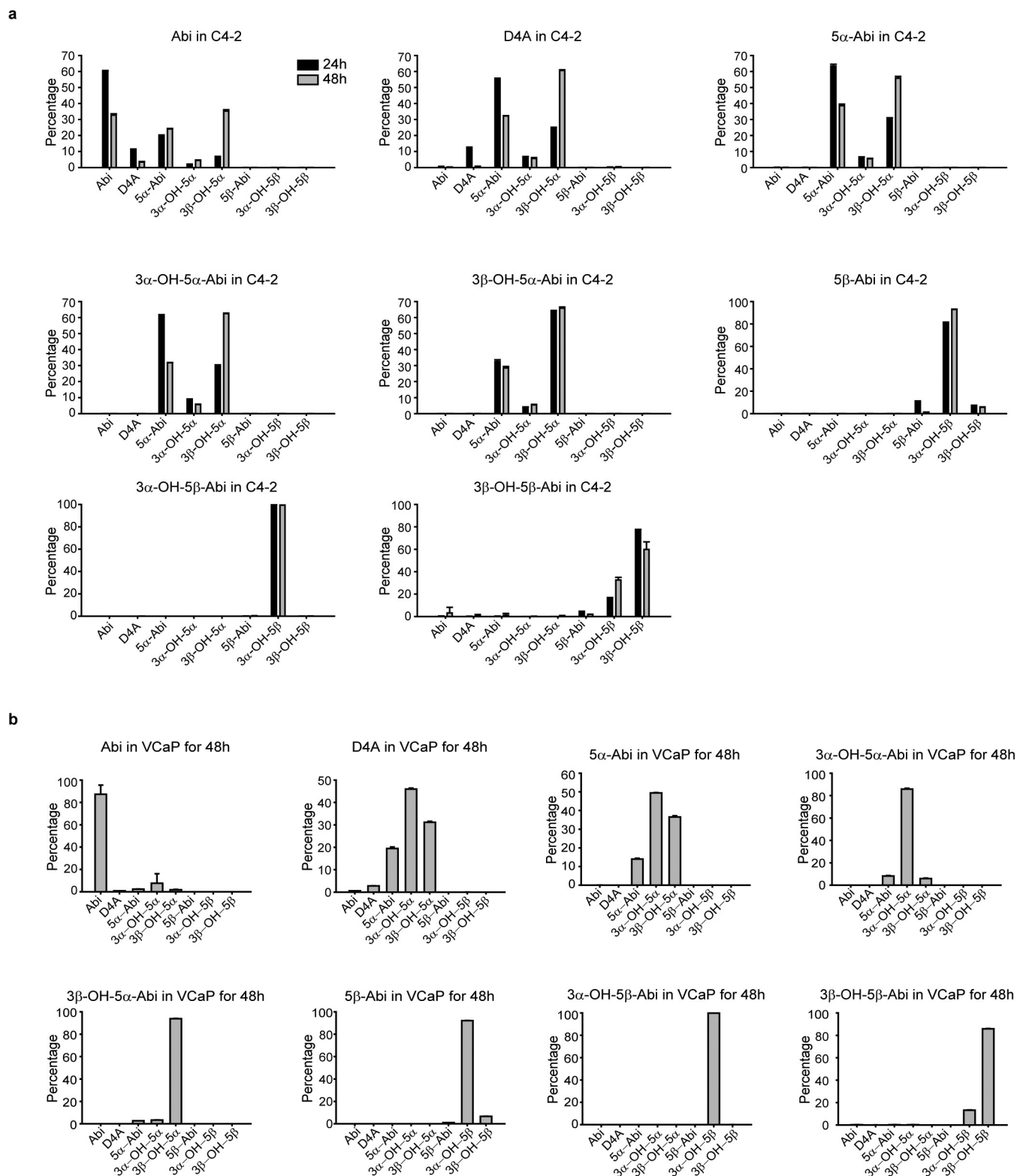
chromatography. Separation of drug metabolites was achieved using a mobile phase consisting of LC-MS grade (Fisher) methanol: acetonitrile: water:formic acid (39:26:34:1) at a flow rate of 0.2 ml min^{-1} , and C18 analytical column; Zorbax Eclipse plus $150 \times 2.1 \text{ mm}$, $3.5 \mu\text{m}$ (Agilent)³².

Chemical Synthesis. See Supplementary Methods.

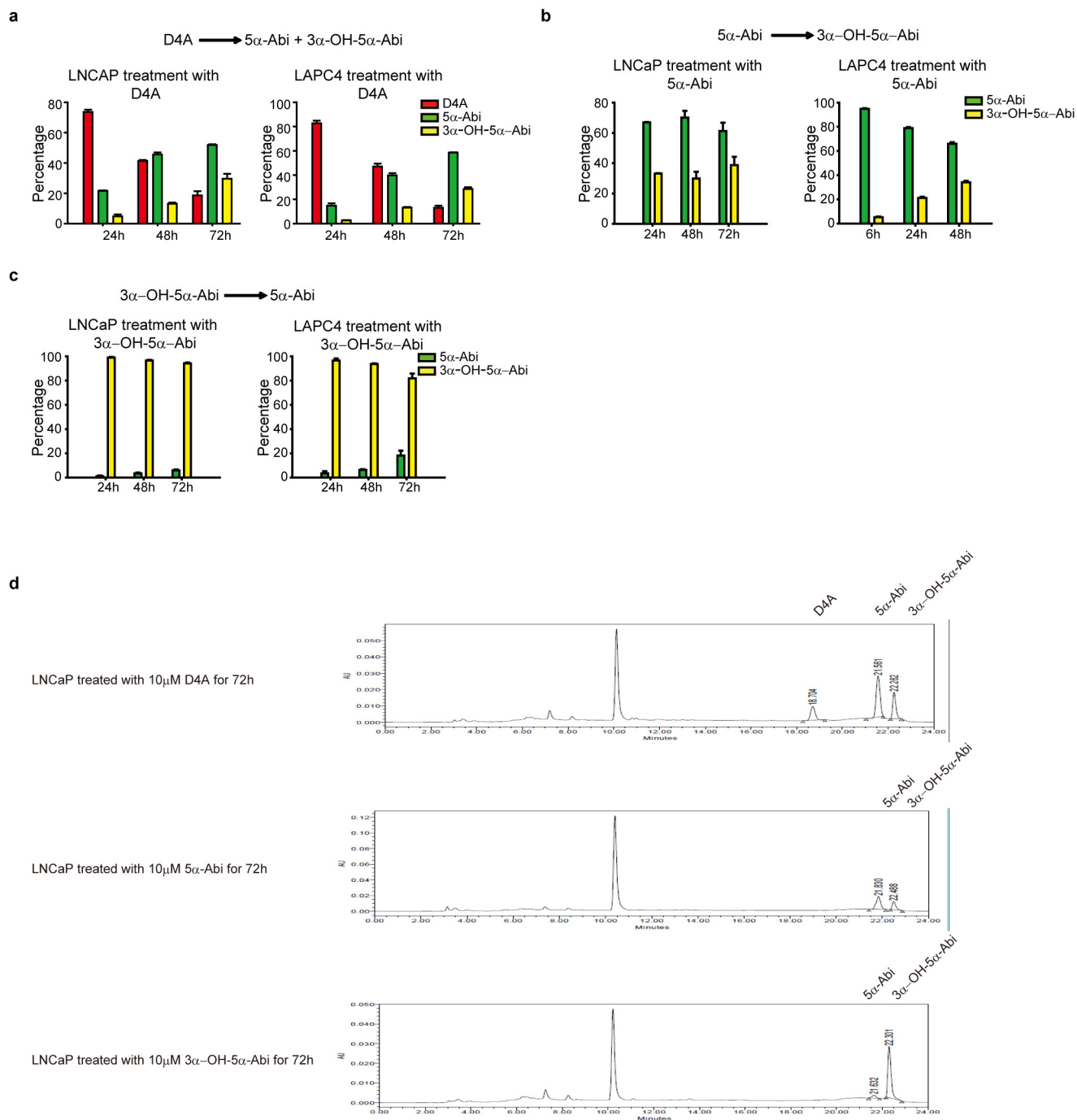
23. Papari-Zareei, M., Brandmaier, A. & Auchus, R. J. Arginine 276 controls the directional preference of AKR1C9 (rat liver 3α -hydroxysteroid dehydrogenase) in human embryonic kidney 293 cells. *Endocrinology* **147**, 1591–1597 (2006).
24. Li, Z. *et al.* Conversion of abiraterone to D4A drives anti-tumour activity in prostate cancer. *Nature* **523**, 347–351 (2015).
25. Liu, L. L. *et al.* Mechanisms of the androgen receptor splicing in prostate cancer cells. *Oncogene* **33**, 3140–3150 (2014).
26. Hörnberg, E. *et al.* Expression of androgen receptor splice variants in prostate cancer bone metastases is associated with castration-resistance and short survival. *PLoS ONE* **6**, e19059 (2011).
27. Deng, W., Wang, Y., Liu, Z., Cheng, H. & Xue, Y. Heml: a toolkit for illustrating heatmaps. *PLoS ONE* **9**, e111988 (2014).
28. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
29. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
30. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
31. Mootha, V. K. *et al.* PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genet.* **34**, 267–273 (2003).
32. Alyamani, *et al.* Development and validation of a novel LC-MS/MS method for simultaneous determination of abiraterone and its seven steroidal metabolites in human serum: innovation in separation of diastereoisomers without use of a chiral column. *J. Steroid Biochem. Mol. Biol.* <http://dx.doi.org/doi:10.1016/j.jsbmb.2016.04.002> (2016).



Extended Data Figure 1 | Synthesis of abiraterone metabolites. a, Synthesis of 5 α -Abi, 3 α -hydroxy-5 α -Abi and 3 β -hydroxy-5 α -Abi. **b,** Synthesis of 5 β -Abi, 3 α -hydroxy-5 β -Abi and 3 β -hydroxy-5 β -Abi. **c,** Synthesis of D4A.

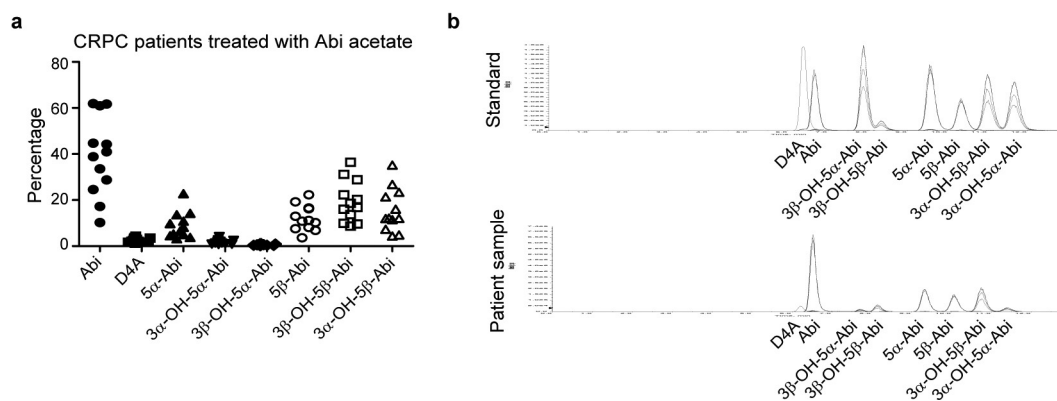


Extended Data Figure 2 | Genesis and interconversion of Abi metabolites. a, C4-2 cells. b, VCaP cells. Cells were treated with abiraterone or the indicated metabolite ($0.1 \mu\text{M}$) for 24 or 48 h and each of the indicated metabolites was detected by LC-MS/MS in triplicate. Error bars represent s.d.

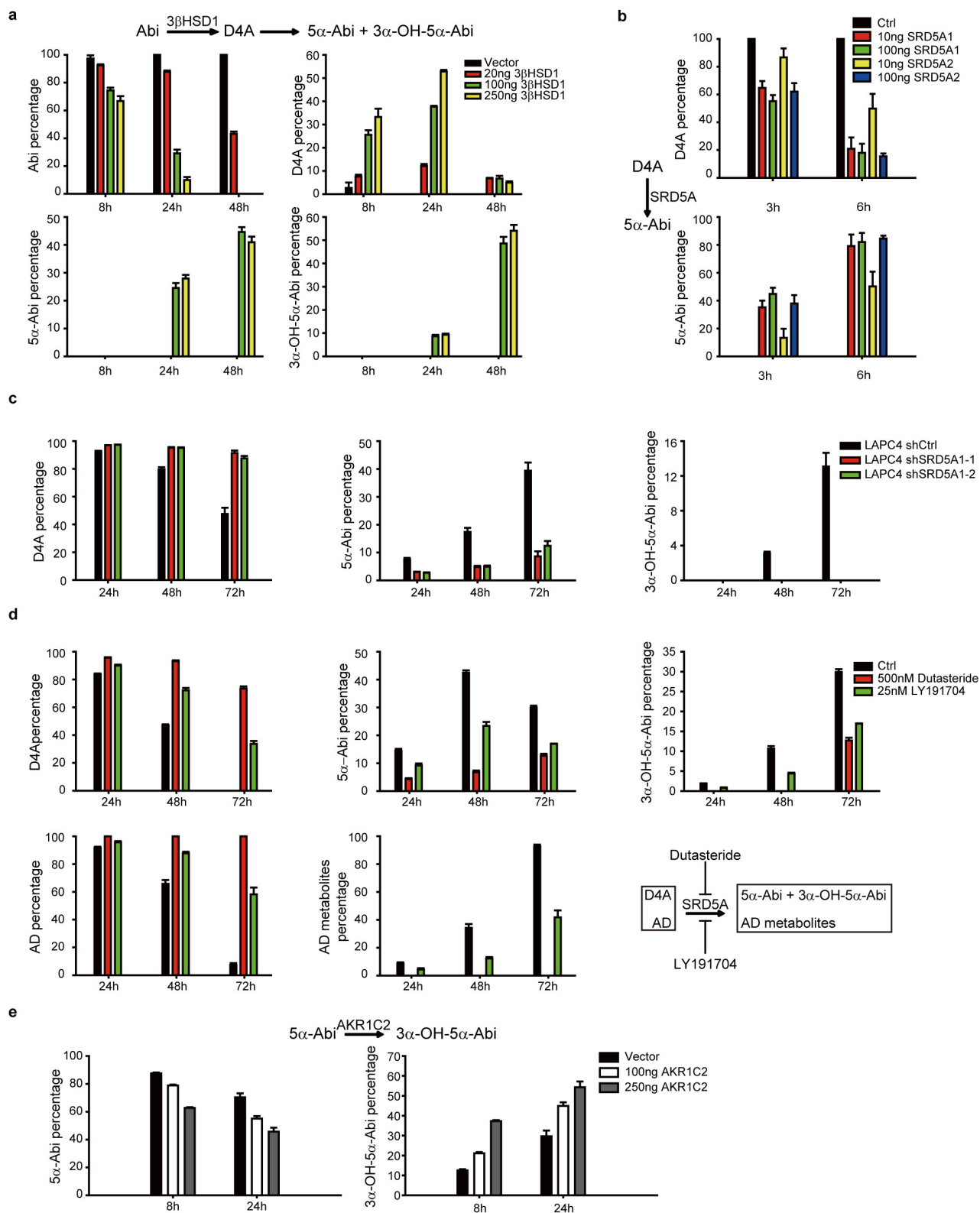


Extended Data Figure 3 | *In vitro* time course formation of 5 α -reduced Abi metabolites. a–c, Conversion from D4A to 5 α -reduced abiraterone metabolites (a), 3-keto reduction of 5 α -Abi to 3 α -OH-5 α -Abi (b), and 3 α -OH-oxidation of 3 α -OH-5 α -Abi to 5 α -Abi (c) is detectable in LNCaP and LAPC4 prostate cancer cell lines. Cells were treated with 10 μ M of the

indicated compounds, metabolites were separated by HPLC and quantified by UV spectroscopy. Experiments were performed in triplicate at least three times and error bars represent s.d. d, Examples of HPLC and UV absorption tracings for incubations of prostate cancer cell lines with D4A, 5 α -Abi and 3 α -OH-5 α -Abi.

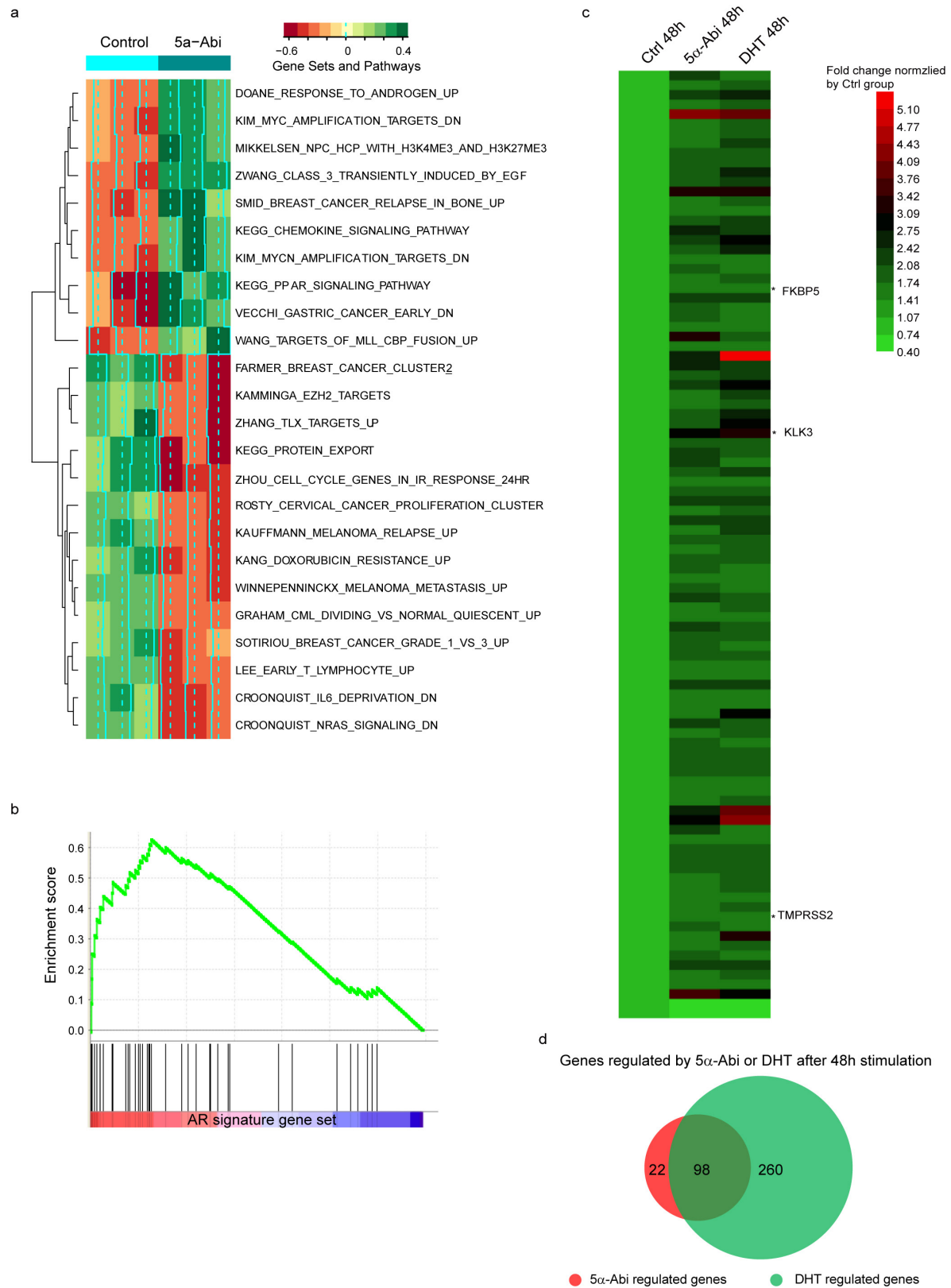


Extended Data Figure 4 | Clinical presence of 5α-reduced and 5β-reduced Abi metabolites in patients treated with Abi acetate. a, Dot plot of Abi and its metabolites expressed as the percentage of the total of Abi and its metabolites. **b,** LC-MS/MS separation of Abi metabolite standards and an example from serum obtained from a patient being treated with Abi.



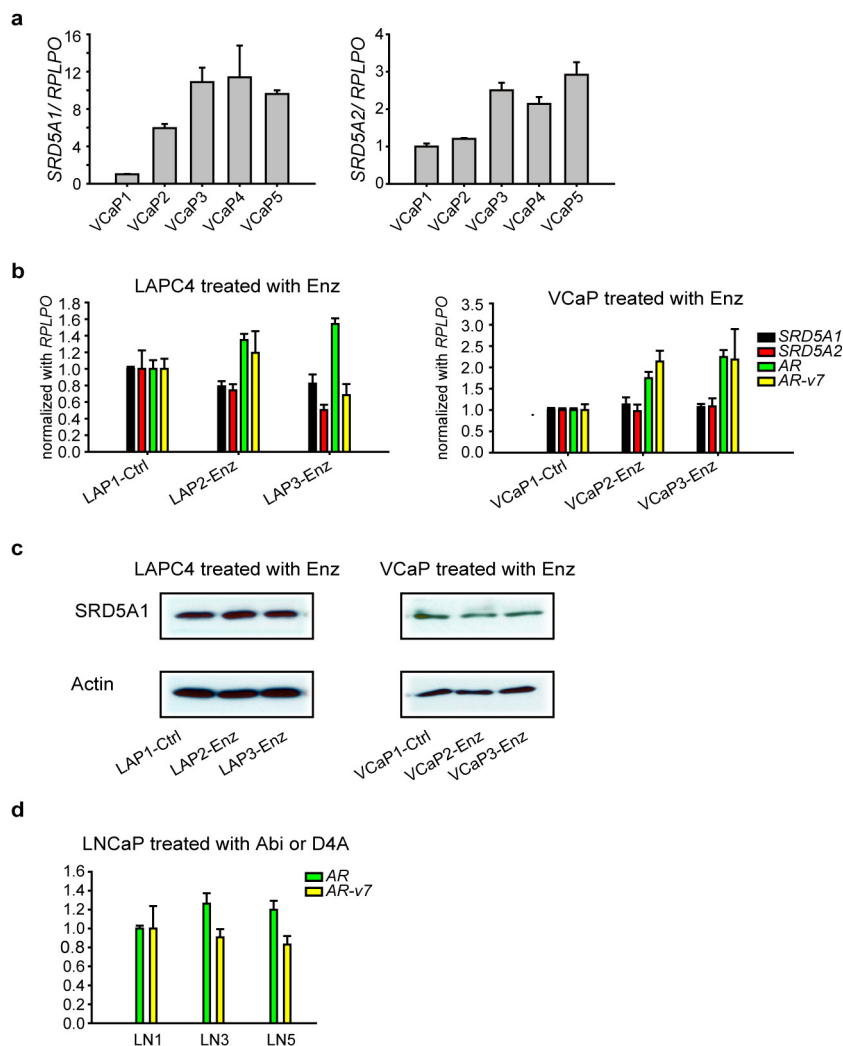
Extended Data Figure 5 | Enzymes involved in the formation of 5 α -reduced Abi metabolites. **a**, $3\beta\text{HSD1}$ catalyses the conversion of Abi to D4A and downstream accumulation of 5 α -Abi and 3 α -OH-5 α -Abi. LAPC4 cells were transiently transfected with the indicated amount of an expression construct encoding $3\beta\text{HSD1}$ or vector control before treatment with Abi. **b**, Conversion of D4A to 5 α -Abi is catalysed by SRD5A1 or SRD5A2. The indicated amounts of SRD5A1, SRD5A2, or empty vector plasmids were transfected into 293T cells, and cells were incubated with D4A for the designated incubation times. **c**, SRD5A1 silencing blocks 5 α -reduction of D4A. LAPC4 cells stably expressing short hairpin (sh) RNAs targeting SRD5A1 or nonsilencing control were treated with D4A

and metabolites for the indicated times. **d**, Pharmacological SRD5A inhibition blocks 5 α -reduction of D4A. LAPC4 cells were treated with D4A and the SRD5A inhibitors dutasteride or LY191704. A parallel control experiment is shown with inhibition of 5 α -reduction of [^3H]androstenedione (AD). **e**, Conversion of 5 α -Abi to 3 α -OH-5 α -Abi is catalysed by AKR1C2. 293T cells were transfected with AKR1C2 or empty vector and treated with 5 α -Abi for the indicated times. For all experiments, metabolites were separated by HPLC and quantified by UV spectroscopy (Abi metabolites) or with a beta-RAM ([^3H]androgens). Error bars represent s.d. All experiments were performed at least three times.



Extended Data Figure 6 | Gene expression profile of stimulation by 5α-Abi and DHT. **a**, Unbiased pathway analysis of 5α-abi-regulated genes. **b**, Gene Set Enrichment Analysis of 5α-abi-regulated genes with the androgen receptor signature gene set. **c**, Gene expression in LAPC4

cells stimulated by 1 μM 5α-Abi or 0.1 nM DHT for 48 h. Regulated genes were determined by detection $P < 0.01$, upregulation > 1.55 or downregulation < 0.5 compared with vehicle control group. **d**, Venn diagram of 5α-abi and DHT-regulated genes.



Extended Data Figure 7 | Transcript expression regulation in the presence of Abi, D4A or enzalutamide. **a**, *SRD5A1* and *SRD5A2* expression in VCaP cells treated with Abi or D4A as indicated in Fig. 3. **b**, *SRD5A1* and *SRD5A2* expression does not change with enzalutamide (Enz) treatment. **c**, *SRD5A1* protein abundance does not change with

enzalutamide treatment in LAPC4 or VCaP cells. **d**, *AR-V7* expression is unchanged in LNCaP cells treated with Abi or D4A as indicated in Fig. 3. Expression was normalized to *RPLPO* and vehicle-treated cells for all comparisons. Error bars represent s.d.

Extended Data Table 1 | Data for each of the 12 patients treated with Abi acetate

Patient	Time from last dose to blood draw	Treatment duration (Month)	Abi (nM)	D4A (nM)	5 α -Abi (nM)	3 α -OH-5 α -Abi (nM)	3 β -OH-5 α -Abi (nM)	5 β -Abi (nM)	3 α -OH-5 β -Abi (nM)	3 β -OH-5 β -Abi (nM)
#1	2 h	32	24.8	3.7	3.9	1.2	0.3	7.0	18.4	26.8
#2	3h 15min	3	611.3	20.7	82.0	8.5	2.8	129.9	48.2	99.5
#3	13h 20min	9	133.5	6.7	43.2	8.4	2.9	65.9	48.3	88.7
#4	2h 30min	26	138.8	3.7	30.6	3.3	1.5	17.1	9.8	19.5
#5	4h 16min	3	26.6	9.5	8.1	1.7	0.4	57.9	60.7	94.6
#6	11h 15min	4	197.0	15.4	43.1	9.3	3.2	49.2	53.0	70.7
#7	6h 20min	7	319.3	14.5	186.9	23.0	5.1	133.6	60.0	79.3
#8	3h 15min	4	6.3	0.8	1.7	0.6	0.2	2.6	9.0	4.3
#9	9h 20min	5	30.8	1.3	2.5	0.7	0.3	1.8	5.7	6.9
#10	8h 10min	4	47.8	8.0	10.4	2.5	1.0	53.3	74.3	79.9
#11	10h 40min	6	102.8	5.2	35.7	10.5	2.8	17.1	30.0	46.6
#12	2h	35	257.5	7.3	31.1	5.9	6.0	63.0	93.5	118.4

Extended Data Table 2 | Genes regulated by 5 α -abi

SYMBOL	5 α -Abi/Ctrl	SYMBOL	5 α -Abi/Ctrl	SYMBOL	5 α -Abi/Ctrl	SYMBOL	5 α -Abi/Ctrl
ABCA1	1.832	FLJ20021*	2.372	LOC440509*	1.558	SEN7*	1.672
ABCC4*†	2.173	FLJ27365	2.034	LOC441763*	1.838	SERHL*	1.613
ABCG1	2.314	FLJ41603*	1.842	LOC554208	1.643	SERHL2*	2.703
ACACB	1.677	FLJ42562*	1.850	LOC643376	2.246	SERPINE2*	2.836
ALDH4A1*	1.712	FOXC1*	1.618	LOC644096	1.760	SGCB*	2.219
ANKRD29*	2.336	FOXN4*	3.227	LOC644584	1.929	SGPP1*	1.580
ANKS3*	1.561	FOXQ1	1.590	LOC646434	1.733	SLAIN1	1.668
ARL15	1.674	FXYD3	1.631	LOC646783*	1.647	SLC16A9*	1.991
BCAR3	1.712	GAL3ST4	2.004	LOC647104*	1.954	SLC45A3*†	1.855
BEND5	1.605	GARNL3*	1.673	LOC651075*	1.754	SNORA58*	1.793
C10orf41	1.566	GDF15*	2.437	LOC728178	1.772	SORD*	1.649
C14orf159	1.831	GHR*	2.458	LOC729799*	1.677	ST3GAL6*	1.605
C1orf116*†	4.181	HDAC11	1.564	LPCAT4*	1.949	STXBP5	1.574
C1orf21*†	1.631	HES6*	1.779	MAFB*	2.184	TBC1D16*	1.597
C1orf74*	1.619	HMGCS2	1.562	MBOAT2	1.576	THAP3	1.811
C5orf41	1.634	HMOX2	1.619	MBP*	1.671	TINF2*	1.687
C6orf105*	1.697	HOMER2*†	2.191	MCOLN2	1.587	TMEM45B	2.023
C7orf38	1.767	HOXC8	1.650	MGC16384	1.886	TMPRSS2*†	1.730
CA12*†	1.821	HS6ST1*	1.671	MIR1974*	1.566	TRIM36*	2.045
CALCB	1.610	HSPA2*	1.666	NCRNA00153	1.552	TRPM8*†	1.601
CAPN2*	1.930	IGSF21*	2.067	NKX3-1*†	2.209	TSPAN33*	1.571
CDC2L2	1.608	JHDM1D	1.621	NOV*	1.844	UHMK1	1.764
CDK6	1.575	KCNK13*	2.066	NRK*	1.761	ULK2*	1.841
CHML	1.553	KCNK17	1.698	NUMBL*	1.902	VWA3A	1.571
COL16A1*	2.017	KLK3*†	3.084	OAZ2*	1.700	WIZ	1.627
CPEB2*	1.899	KLK4*	1.802	ODF3L2*	1.650	YPEL1*	2.116
CPT1C	1.604	KRT126P*	2.348	PALMD*	2.252	ZFX*	1.717
CUEDC1	1.774	LEPROT	1.831	PCDH20	1.969	ZNF264	1.553
CYP1B1	4.321	LOC10008589*	2.206	PIGL*	1.671	ZNF30*	1.577
DDIT4L*	3.243	LOC100129674	1.586	PM20D1	1.555	ZNF350	1.751
DIP2C*	1.684	LOC100130123*	1.893	PNPLA7*	1.627	ZNF414	1.899
DIS3L2	1.696	LOC100132394*	1.700	PP8961*	1.554	ZPLD1*	3.579
EDN1*	1.575	LOC100132564*	1.894	PPFIBP2*	2.087	B3Gn-T6*	0.494
EMP1*	1.760	LOC100133099	1.585	PPP1R3E*	1.563	CAMK1G	0.423
ENDOD1*†	2.467	LOC100133565*	2.217	PRICKLE1*	1.840	CGA	0.476
EPR1*	2.130	LOC145837*	1.680	PTPRR*	1.750	FANCB	0.445
ETS2*	1.998	LOC388681	1.677	PXK*	1.880	LOC100130775*	0.473
FAM129A*	1.576	LOC389286*	2.196	REPS2*	1.807	LOC347376	0.495
FAM134B*	2.035	LOC389901	1.566	RHOA	1.584	LOC440063	0.445
FAM46B*	1.688	LOC400214	1.552	RNF144A	1.563	MALL	0.495
FKBP5*†	1.640	LOC440122	1.617	SASH1*	1.560	TTC7B	0.477

*: co-regulated by DHT.

†: reported bona fide AR target gene.

Extended Data Table 3 | Concentrations of Abi and its metabolites in a phase II clinical trial (NCT01393730)

Patient		#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16
Abi	cycle 3	439.46	64.74	210.65	447.29	80.78	195.53	117.58	79.58	135.63	200.93	190.73	320.19	66.11	177.31	131.86	205.48
	cycle 4	1049.04	136.18	95.97	400.55	1399.54	288.80	214.26	207.19	148.05	393.35	219.20	569.72	173.93	203.84	178.06	289.94
	cycle 7	1767.26	130.53	268.79	883.07	112.89	174.05	310.03	127.05	204.86	52.92	220.11	184.44	69.55	215.29	59.88	112.91
D4A	cycle 3	14.83	2.78	11.80	8.89	7.24	12.32	17.23	4.01	10.62	19.11	4.95	11.78	5.23	10.79	11.83	5.91
	cycle 4	34.24	8.83	7.75	13.86	34.49	28.84	23.70	13.01	11.01	16.71	11.91	32.84	12.35	13.09	17.91	10.73
	cycle 7	69.00	8.32	13.64	36.14	15.61	18.63	17.47	7.96	15.60	4.98	10.89	11.01	7.04	32.40	10.64	6.71
3-keto-5 α -Abi	cycle 3	44.53	13.29	26.07	16.96	17.46	23.66	22.70	19.74	29.27	38.73	25.28	34.24	21.78	27.32	14.79	36.98
	cycle 4	1.73	4.28	1.78	0.80	3.53	1.53	2.89	0.94	2.04	1.95	0.91	1.64	8.36	3.44	8.44	2.80
	cycle 7	4.34	0.49	0.71	8.97	9.43	0.81	1.41	0.61	13.65	0.56	0.86	0.80	4.86	11.30	6.95	5.05
3 α -OH-5 α -Abi	cycle 3	15.81	4.45	4.04	3.80	4.53	5.65	4.12	4.86	8.80	9.57	5.56	8.82	5.52	6.16	2.69	3.27
	cycle 4	0.22	0.01	0.18	0.14	0.77	0.35	0.53	0.16	0.52	0.51	0.08	0.01	1.85	0.48	1.91	0.28
	cycle 7	0.84	0.00	0.00	2.45	1.35	0.05	0.00	0.00	2.81	0.08	0.00	0.12	0.98	1.77	1.92	0.49
3 β -OH-5 α -Abi	cycle 3	5.15	1.31	2.31	2.26	3.25	2.00	1.21	1.63	2.75	3.43	1.55	2.87	2.54	5.81	3.89	1.06
	cycle 4	0.31	0.03	0.60	0.21	2.16	0.09	0.31	0.06	0.32	0.20	0.09	0.17	1.17	1.93	3.83	0.16
	cycle 7	0.59	0.01	0.13	4.04	2.32	0.04	0.17	0.07	3.22	0.05	0.12	0.12	0.50	5.59	2.69	0.31
3-keto-5 β -Abi	cycle 3	15.63	6.24	93.39	31.14	55.55	12.21	39.43	12.83	52.75	24.97	23.05	25.00	18.55	57.78	78.37	73.90
	cycle 4	53.05	20.41	41.95	43.95	65.31	24.55	51.38	58.97	24.22	63.34	26.42	36.53	28.47	60.61	73.60	127.48
	cycle 7	141.23	17.17	42.09	75.30	101.40	25.41	24.44	24.65	47.59	20.24	34.29	18.47	29.70	205.11	56.29	102.62
3 α -OH-5 β -Abi	cycle 3	41.95	14.63	86.05	44.46	70.25	19.57	37.12	21.94	86.71	31.06	31.93	42.80	33.90	98.26	76.44	28.59
	cycle 4	42.31	36.79	55.14	27.67	34.26	34.66	70.21	53.53	75.32	44.13	31.55	76.59	42.91	39.17	68.02	36.96
	cycle 7	110.85	29.24	63.35	99.70	48.28	35.00	37.37	38.95	57.38	24.67	46.16	34.37	33.17	142.07	68.89	37.41
3 β -OH-5 β -Abi	cycle 3	35.68	18.41	121.85	96.16	129.09	23.92	33.38	33.95	103.67	59.47	41.23	65.71	31.22	216.92	127.26	48.69
	cycle 4	36.33	40.71	115.43	54.54	50.89	35.06	72.74	78.52	82.78	77.29	35.08	83.37	39.21	125.80	114.74	71.33
	cycle 7	95.93	26.97	80.34	180.42	114.67	40.48	41.48	52.28	73.76	40.39	51.22	47.96	42.08	387.83	106.73	82.47

Access of protective antiviral antibody to neuronal tissues requires CD4 T-cell help

Norifumi Iijima¹ & Akiko Iwasaki¹

Circulating antibodies can access most tissues to mediate surveillance and elimination of invading pathogens. Immunoprivileged tissues such as the brain and the peripheral nervous system are shielded from plasma proteins by the blood–brain barrier¹ and blood–nerve barrier², respectively. Yet, circulating antibodies must somehow gain access to these tissues to mediate their antimicrobial functions. Here we examine the mechanism by which antibodies gain access to neuronal tissues to control infection. Using a mouse model of genital herpes infection, we demonstrate that both antibodies and CD4 T cells are required to protect the host after immunization at a distal site. We show that memory CD4 T cells migrate to the dorsal root ganglia and spinal cord in response to infection with herpes simplex virus type 2. Once inside these neuronal tissues, CD4 T cells secrete interferon- γ and mediate local increase in vascular permeability, enabling antibody access for viral control. A similar requirement for CD4 T cells for antibody access to the brain is observed after intranasal challenge with vesicular stomatitis virus. Our results reveal a previously unappreciated role of CD4 T cells in mobilizing antibodies to the peripheral sites of infection where they help to limit viral spread.

To investigate the mechanism of antibody-mediated protection within the barrier-protected tissues, we used a mouse model of genital herpes infection. Herpes simplex virus type 2 (HSV-2) enters the host through the mucosal epithelia, and infects the innervating neurons in the dorsal root ganglia (DRG) to establish latency^{3,4}. Vaginal immunization by an attenuated HSV-2 with deletion of the thymidine kinase gene (TK⁻ HSV-2) provides complete protection from lethal disease following genital challenge with wild-type (WT) HSV-2 (ref. 5) by establishing tissue-resident memory T cells (T_{RM})⁶. In vaginally immunized mice, interferon (IFN)- γ secretion by CD4 T cells, but not antibodies, are required for protection^{7,8}. In contrast, distal immunization with the same virus fails to establish T_{RM} and provides only partial protection⁶. Nevertheless, of the distal immunization routes tested, intranasal immunization with TK⁻ HSV-2 provided the most robust protection against intravaginal challenge with WT HSV-2, whereas intraperitoneal immunization provided the least protection (Fig. 1a–d)^{9,10}. As shown previously⁶, intranasal immunization did not establish T_{RM} in the genital mucosa (Extended Data Fig. 1a, b), despite generating a comparable circulating memory T-cell pool (Extended Data Fig. 1c, d). After vaginal HSV-2 challenge, mice that were immunized intranasally with TK⁻ HSV-2 were unable to control viral replication within the vaginal

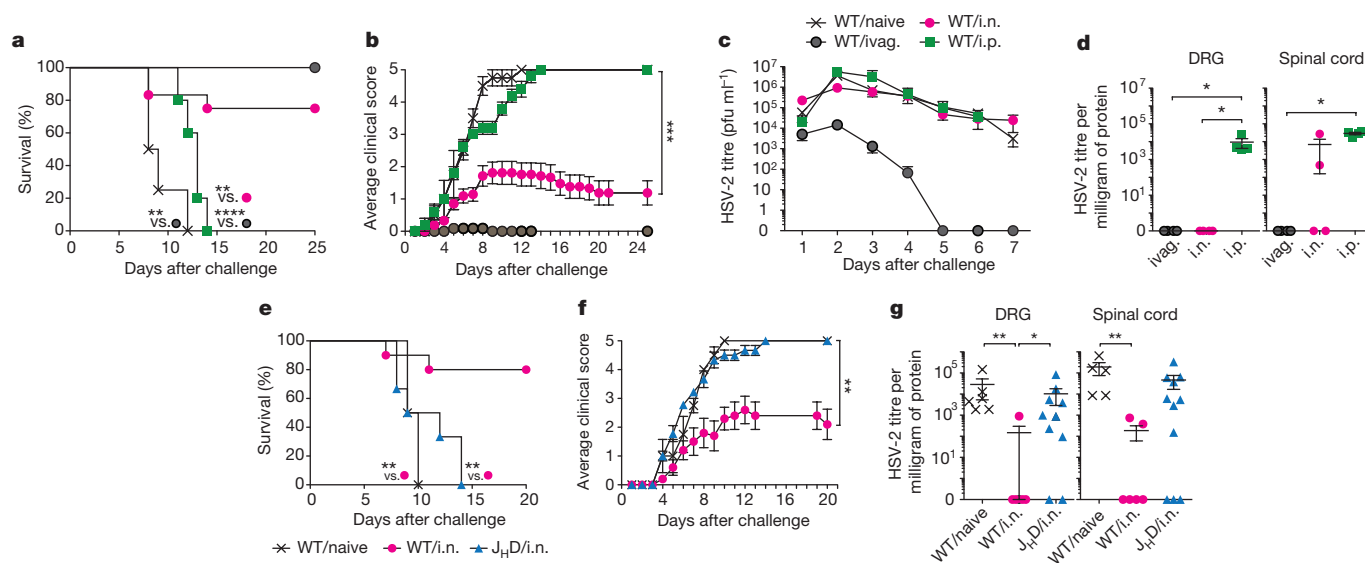


Figure 1 | Intranasal immunization confers B-cell-dependent neuron protection following genital HSV-2 challenge. a–d, C57BL/6 mice were immunized with TK⁻ HSV-2 (10^5 plaque-forming units (p.f.u.)) via intranasal (i.n.; $n = 12$), intraperitoneal (i.p.; $n = 5$) or intravaginal (ivag.; $n = 11$) routes. Five to six weeks later, these mice and naive mice ($n = 4$) were challenged with a lethal dose of WT HSV-2 (10^4 p.f.u.). Mortality (a), clinical score (b) and virus titre in vaginal wash (c) were measured on indicated days after challenge. d, Six days after challenge, virus titre in tissue homogenates including DRG and spinal cord was measured.

e–g, BALB/c mice ($n = 10$) or B-cell-deficient JHD mice ($n = 6$) were immunized intranasally with TK⁻ HSV-2 (5×10^4 p.f.u.). Six weeks later, these mice and naive mice ($n = 4$) were challenged with lethal WT HSV-2 (10^5 p.f.u.). Mortality (e) and clinical score (f) were measured. g, Six days after challenge, virus titre in tissue homogenates including DRG and spinal cord was measured by plaque assay. Data are mean \pm s.e.m. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; **** $P < 0.0001$ (two-tailed unpaired Student's *t*-test).

¹Howard Hughes Medical Institute, Department of Immunobiology, Yale School of Medicine, New Haven, Connecticut 06520, USA.

mucosa (Fig. 1c), but had significantly reduced viral replication in the innervating neurons of the DRG (Fig. 1d). Notably, we found that protection conferred by intranasal immunization required B cells, as JH mice (deficient in B cells) were not protected by intranasal immunization (Fig. 1e–g). In the absence of B cells, intranasal immunization was unable to control viral replication in the DRG and spinal cord (Fig. 1g).

In mice immunized intranasally with TK[−] HSV-2, no evidence of vaccine virus in the DRG or the spinal cord was found (Extended Data Fig. 1e). Moreover, the intranasal route of immunization was not unique in conferring protective response, as parabiotic mice sharing circulation with intravaginally immunized partners were also partly protected from vaginal challenge with WT HSV-2 in the absence of T_{RM}⁶ (Extended Data Fig. 1f–h). We found that the B cells in the immunized partners were required to confer protection in the naive conjoined mice, as partners of immunized μ MT mice were unprotected (Extended Data Fig. 1f–h). Moreover, antigen-specific B cells were required to confer protection, as intravaginally immunized partners whose B cells bore an irrelevant B cell receptor (against hen egg lysozyme (HEL)) were unable to confer protection in the conjoined naive partner (Extended Data Fig. 1f–h). As observed for the intranasal immunization, viral control conferred by the immunized parabiotic partner was not observed in the vaginal mucosa (Extended Data Fig. 1h), suggesting that protection occurs in the innervating neurons.

Next, we investigated the basis for superior protection by antibodies following different routes of immunization. Intravaginal, intranasal and intraperitoneal routes of immunization with TK[−] HSV-2 results in comparable circulating CD4 T-cell memory responses⁶. While no differences were seen for other isotypes, the intranasal and intravaginal routes of immunization were superior to intraperitoneal route in generating higher levels of systemic HSV-2-specific immunoglobulin-G (IgG)2b and IgG2c responses (Extended Data Fig. 2). These results indicated that higher levels of circulating virus-specific IgG2b and IgG2c correlate with protection against vaginal HSV-2 challenge.

We next examined how antibody access to the DRG and spinal cord is mediated. Even though the peripheral nervous tissues are protected from antibody diffusion through the blood–nerve barrier, it was formally possible that secretion of antibody into the tissue occurs through transport of serum antibody by the neonatal Fc receptor for IgG (FcRn)¹¹ expressed on the endothelial cells within the infected tissues. However, we found that mice deficient in FcRn immunized intranasally with TK[−] HSV-2 were equally protected as the WT counterpart from vaginal HSV-2 infection (Fig. 2a, b). Thus, circulating HSV-2-specific antibodies are somehow mobilized to the neuronal tissues following local viral infection in an FcRn-independent manner, and are required for protection of the host.

If circulating antibodies are sufficient, passive transfer of HSV-2-specific antibodies alone should be able to protect the host. However, we and others^{12,13} have found that intravenous injection of HSV-2-specific antibodies alone fails to protect naive mice against HSV-2 challenge (Fig. 2c, d). In contrast, consistent with a previous study¹³, we found that B-cell-deficient μ MT mice immunized intranasally with TK[−] HSV-2 and given systemic administration of HSV-2-specific antiserum were protected (Fig. 2c, d). Thus, these results demonstrate that it is the secreted antibodies, and not B cells themselves, in concert with non-B-cell immune cells, probably T cells induced by immunization, that seem to be required for protection. To test this possibility, we depleted CD4 T cells from mice previously immunized intranasally just before intravaginal HSV-2 challenge. In this setting, differentiation of B cells and antibody responses were allowed to occur fully in the presence of CD4 T-cell help for 6 weeks. Mice acutely depleted of CD4 T cells succumbed to challenge with HSV-2 (Fig. 2e, f), whereas depletion of CD8 T cells and natural killer (NK) cells had no effect⁹. Moreover, neutralization of IFN- γ before challenge, or genetic deficiency in IFN- γ R, also rendered intranasally immunized mice more susceptible to intravaginal HSV-2 challenge (Fig. 2e, f). Of note, depletion of CD4 T cells from intranasally immunized mice just before the

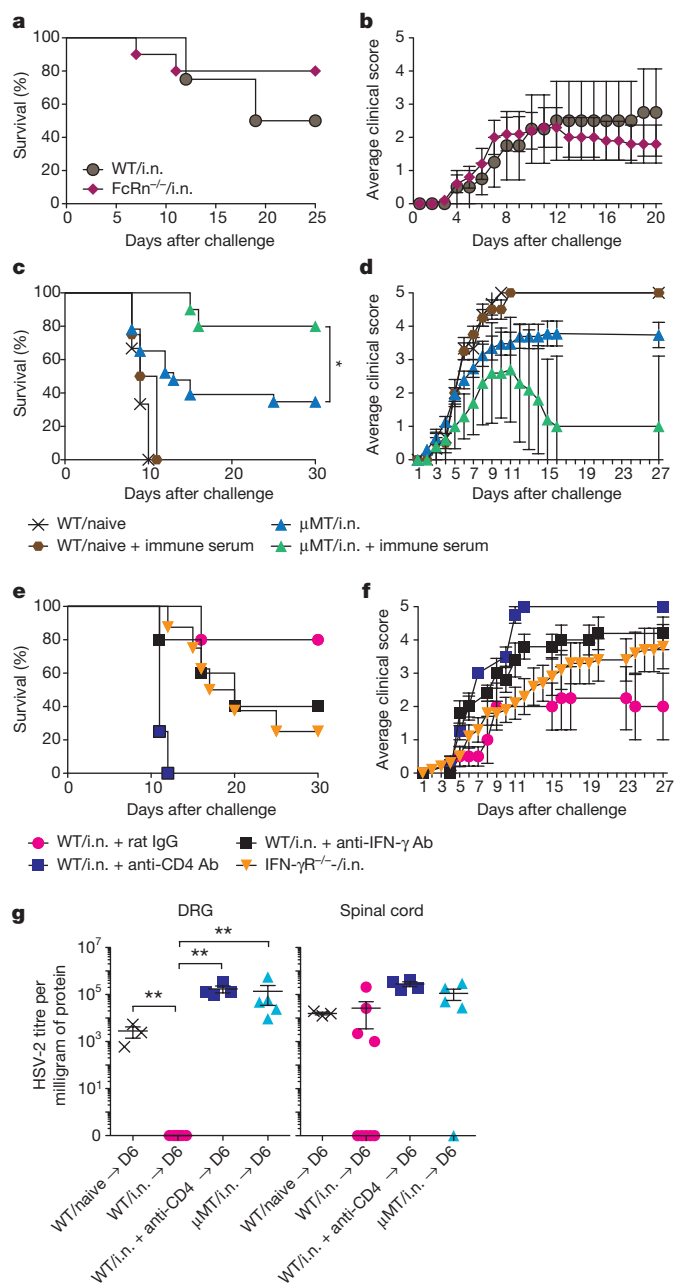


Figure 2 | Antibody-mediated neuroprotection depends on CD4 T cells but not on FcRn-mediated transport. **a, b**, C57BL/6 (WT) mice ($n = 4$) and FcRn^{−/−} ($n = 10$) mice were immunized intranasally with TK[−] HSV-2 (10^5 p.f.u.), and 6 weeks later challenged with a lethal dose of WT HSV-2 (10^4 p.f.u.). Mortality (**a**) and clinical score (**b**) were measured. **c, d**, μ MT mice were immunized with TK[−] HSV-2 (10^5 p.f.u.) intranasally. Five to 6 weeks later, naive mice ($n = 3$), naive mice receiving immune serum intravenously ($n = 4$), μ MT mice ($n = 23$) and μ MT mice receiving immune serum intravenously ($n = 10$) were challenged with a lethal dose of WT HSV-2, and mortality (**c**) and clinical score (**d**) were assessed. Immune serum prepared from mice immunized 4 weeks previously with TK[−] HSV-2 (200μ l per mouse) was injected 3 h before challenge, and 3 and 6 days after challenge. **e, f**, WT C57BL/6 mice ($n = 5$) and IFN- γ R^{−/−} mice ($n = 8$) immunized intranasally with TK[−] HSV-2 (10^5 p.f.u.) 6 weeks previously were challenged with a lethal dose of WT HSV-2, and mortality (**e**) and clinical score (**f**) were assessed. Depletion of CD4 T cells ($n = 4$) or neutralization of IFN- γ ($n = 5$) was performed on days −4, and −1, 2 and 4 after challenge by intravenous injection of anti-CD4 (GK1.5) or anti-IFN- γ (XMG1.2), respectively. **g**, Six days after challenge, virus titre in tissue homogenates including DRG and spinal cord was measured by plaque assay (**e**). Data are mean \pm s.e.m. * $P < 0.05$; ** $P < 0.01$ (two-tailed unpaired Student's *t*-test).

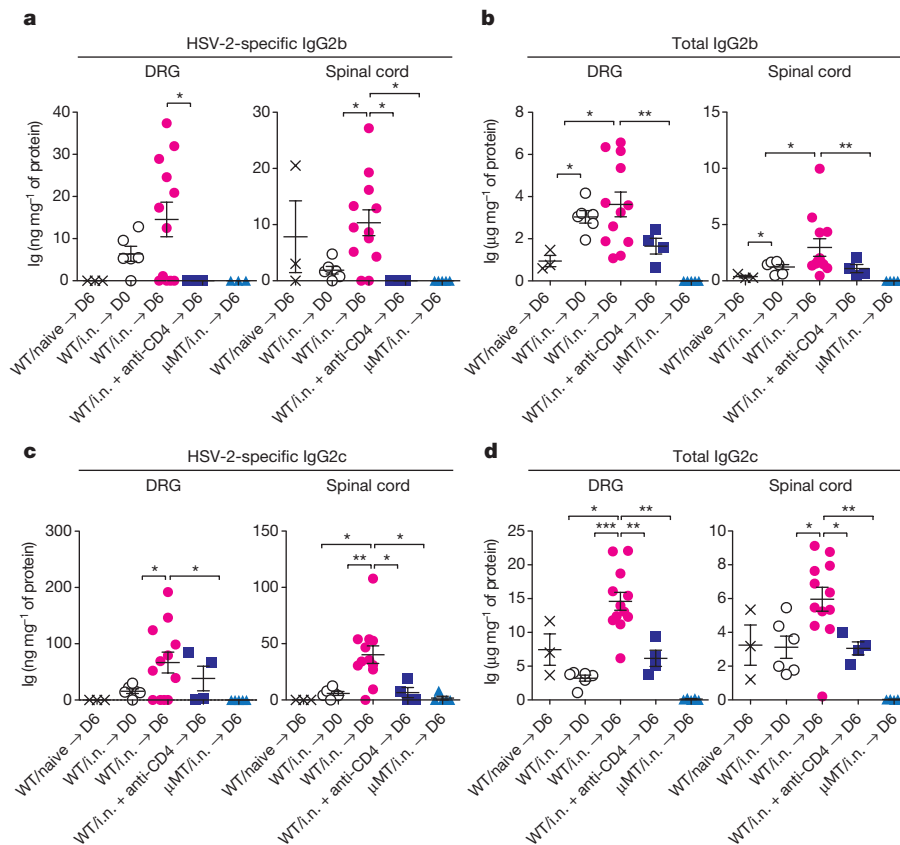


Figure 3 | Memory CD4⁺ T cells are required for antibody access to neuronal tissues. **a–d**, Naive WT mice or WT and μ MT mice intranasally immunized with TK⁻ HSV-2 (10^5 p.f.u.) 6 weeks earlier were challenged with a lethal dose of WT HSV-2 intravaginally. Six days after the challenge, after extensive perfusion, HSV-2-specific (**a**, **c**) and total Ig (**b**, **d**) levels

in tissue homogenates of DRG and spinal cord were analysed by ELISA. To deplete CD4 T cells, CD4-specific antibody was injected on days -4, and -1, 2 and 4 days after challenge. Data are mean \pm s.e.m. * P < 0.05; ** P < 0.01; *** P < 0.001 (two-tailed unpaired Student's t -test).

viral challenge rendered mice incapable of viral control in the DRG, to a similar extent as the immunized B-cell-deficient μ MT mice (Fig. 2g). We observed that intranasal immunization conferred near-complete protection from HSV-2 in the DRG but variable protection in the spinal cord (Figs 1d and 2g). Because HSV-2 can differentially seed the DRG and spinal cord through sensory neurons and autonomic neurons¹⁴, these data suggest that the efficacy of antibody-mediated protection may depend on the route of viral entry. Further, these results indicate that circulating antibodies, CD4 T cells and IFN- γ collectively mediate neuroprotection against HSV-2.

Given that antibody-mediated protection occurs at the level of the innervating neurons and not within the vagina (Fig. 1c and Extended Data Fig. 1h), we hypothesized that CD4 T cells might control delivery of antibodies to the tissue parenchyma through secretion of IFN- γ . We detected only low levels of virus-specific and total antibodies in the DRG or spinal cord at steady state in immunized mice (Fig. 3; WT/intranasally \rightarrow D0), and undetectable levels of antibodies in these tissues in previously unimmunized mice 6 days after an acute infection with HSV-2 (Fig. 3; WT/naive \rightarrow D6). However, in mice immunized intranasally with TK⁻ HSV-2 6 weeks earlier, increase in the levels of antibodies was detected 6 days after intravaginal HSV-2 challenge within the DRG and in the spinal cord (Fig. 3; WT/intranasally \rightarrow D6). Moreover, CD4 T cells were required for access of virus-specific antibodies to the restricted tissue such as the DRG, as depletion of CD4 T cells completely diminished antibody levels in this tissue and spinal cord (Fig. 3d; WT/intranasally + anti-CD4 \rightarrow D6). Further, similar requirement for CD4 T cells (Fig. 3b, d) and IFN- γ (Extended Data Fig. 3) was found for diffusion of total IgG2b and IgG2c isotypes into the DRG, indicating that the delivery mechanism does not

discriminate virus-specificity of the antibodies. In contrast to the neuronal tissues, acute depletion of CD4 or IFN- γ blockade once antibody responses were established had no significant impact on the serum levels of anti-HSV-2 or total antibodies (Extended Data Fig. 4a, b). To examine whether antigen-specific memory CD4 T cells were required to mediate antibody access to the neuronal tissues, mice were primed intranasally with an unrelated virus, influenza A virus and, 4 weeks later, were challenged with HSV-2 intravaginally. In contrast to mice harbouring cognate memory CD4 T cells, antibody access to neuronal tissues following intravaginal HSV-2 challenge was not observed in mice that had irrelevant memory CD4 T cells (against influenza A virus) (Extended Data Fig. 5). These data indicate that antigen-specific memory CD4 T cells are required for antibody access to the neuronal tissues.

We hypothesized that memory CD4 T cell might enter the barrier-protected tissues and mobilize antibody access through local secretion of IFN- γ . In support of this idea, we found that IFN- γ -secreting HSV-2-specific CD4 T cells entered the DRG and spinal cord around 6 days after genital HSV-2 challenge in mice that received intranasal immunization 6 weeks previously (Fig. 4a, b; WT/intranasally \rightarrow D6). Some increase in innate leukocytes bearing CD11b, Ly6G or MHCII was observed in DRG and spinal cord 6 days after challenge (Extended Data Fig. 6a). IFN- γ secretion was confined to the memory CD4 T-cell population within the DRG (Fig. 4a). Moreover, entry of effector CD4 T cells to the DRG and spinal cord at 6 days after primary vaginal HSV-2 infection was much less efficient than their memory counterpart (Fig. 4a, b; WT/naive \rightarrow D6), suggesting the intrinsic ability of T cells to migrate into these neuronal tissues is enhanced with memory development.

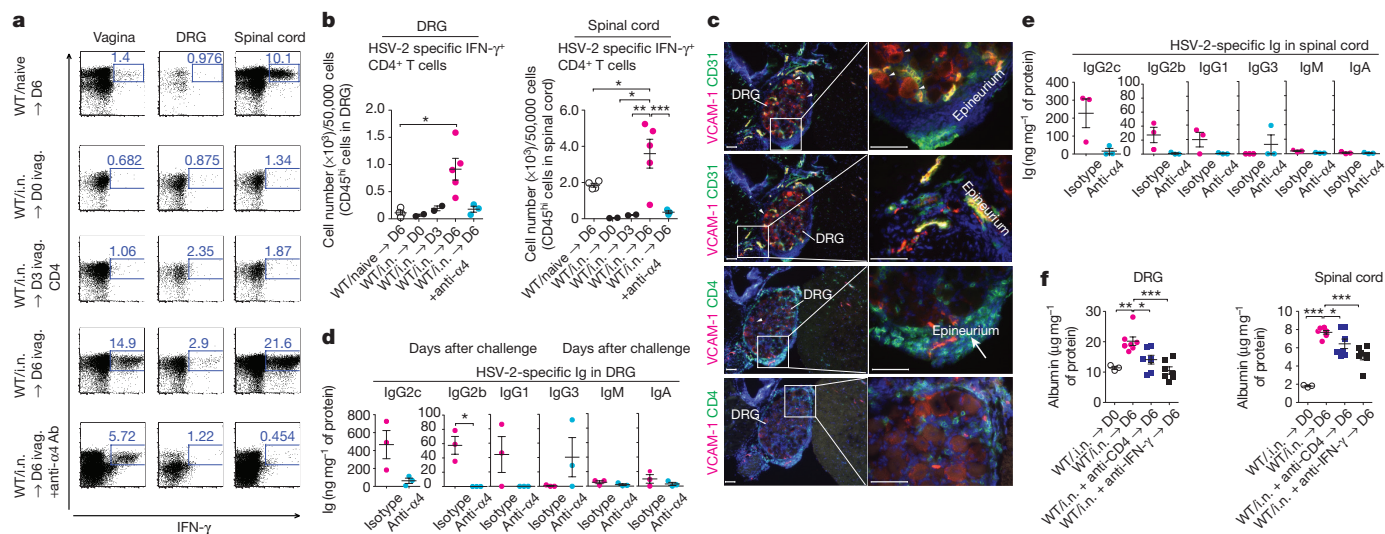


Figure 4 | α 4-Integrin-dependent recruitment of memory CD4⁺ T cells is required for antibody access to neuronal tissues. WT mice immunized intranasally with TK⁻ HSV-2 6 weeks earlier were challenged with a lethal dose of WT HSV-2. Neutralization of α 4-integrin was performed on days 2 and 4 after challenge by intravenous injection of anti- α 4 integrin (CD49d) antibody. **a**, Six days after challenge, after extensive perfusion, HSV-2-specific IFN- γ ⁺ CD4⁺ T cells in DRG and spinal cord were detected by flow cytometry. **b**, The number of IFN- γ -secreting CD4⁺ T cells among 50,000 cells of CD45⁺ leukocytes in DRG and spinal cord is depicted. Data are mean \pm s.e.m. * P < 0.05; ** P < 0.01; *** P < 0.001 (two-tailed unpaired Student's *t*-test). **c**, Frozen sections of DRG were stained with antibodies against CD4, VCAM-1 or CD31. Nuclei are

depicted by 4',6-diamidino-2-phenylindole (DAPI) stain (blue). Images were captured using a $\times 10$ or $\times 40$ objective lens. Scale bars, 100 μ m. Arrowhead indicates VCAM-1⁺ cells in parenchyma of DRG. Data are representative of at least three similar experiments. HSV-2-specific antibodies in the DRG (**d**) and spinal cord (**e**) were analysed by ELISA. Data are mean \pm s.e.m. * P < 0.05 (two-tailed paired Student's *t*-test). Albumin level in tissue homogenates was analysed by ELISA (**f**). Depletion of CD4 T cells or neutralization of IFN- γ was performed on days -4, and -1, 2 and 4 days after challenge by intravenous injection of anti-CD4 (GK1.5) or anti-IFN- γ (XMG1.2), respectively. Data are mean \pm s.e.m. * P < 0.05; ** P < 0.01; *** P < 0.001 (two-tailed paired Student's *t*-test).

Interaction of α 4 β 1 (or VLA4) and VCAM-1 contributes to T-cell recruitment across the blood–brain barrier¹⁵. Memory CD4 T cells generated against HSV-2 expresses CD49d which is the integrin α 4 subunit⁶. We found that the entry of memory CD4 T cells into the nervous tissue was strictly dependent on α 4 integrin, as antibody blockade of α 4 prevented their entry into the DRG and spinal cord (Fig. 4a, b). We observed expression of ligand for α 4 β 1, VCAM-1, in the endothelium of DRG and spinal cord in immune-challenged mice (Fig. 4c and Extended Data Fig. 6b). Further, analysis of tissue sections revealed that the CD4 T cells were found in the parenchyma of the DRG and spinal cord, as well as within their epineurium and meninges, but not within the vasculature (Fig. 4c and Extended Data Fig. 6a, b). Notably, many CD4 T cells were found adjacent to the cell body of neurons within the DRG. Some VCAM-1 staining was found in the cytosol of neuronal cell bodies (arrowhead Fig. 4c). Additionally, intravascular staining¹⁶ with antibody to CD90.2 revealed that the vast majority of the CD4 T cells in the DRG and spinal cord are sequestered from circulation (Extended Data Fig. 7a, b). Thus, CD4 T cells recruited to the neuronal tissues access the parenchyma of the DRG and spinal cord. Notably, α 4 integrin blockade of CD4 T-cell recruitment resulted in diminished access of virus-specific antibody to the DRG and spinal cord (Fig. 4d, e), with no effect on blood levels of virus-specific antibody (Extended Data Fig. 4c) or the total antibody levels of various isotypes in circulation (Extended Data Fig. 4d). Collectively, these data indicate that memory CD4 T cells enter the neuronal tissue and secrete IFN- γ to promote antibody access to the DRG and spinal cord.

How might IFN- γ secreted by CD4 T cells enable circulating antibody to access the neuronal tissues? It is well known that IFN- γ acts on the endothelial cells to remodel tight junctions and increase permeability¹⁷. We observed that recombinant IFN- γ injected intravaginally was sufficient to enable antibody access to the vaginal lumen, suggesting that IFN- γ is sufficient to induce both vascular and epithelial permeability in peripheral tissues (Extended Data Fig. 8a) and to enhance VCAM-1 expression on endothelial cells (Extended

Data Fig. 8b). To assess whether antibody access to the neuronal tissues mediated by CD4 T cells and IFN- γ is through increased vascular permeability, we measured release of blood albumin into the neuronal tissue following genital HSV-2 challenge in intranasally immunized mice. Notably, we observed that vascular permeability occurred in the DRG and spinal cord in a CD4 T-cell- and IFN- γ -dependent manner, as measured by leakage of blood albumin to the neuronal tissues by ELISA and immunohistochemical analysis (Fig. 4f and Extended Data Fig. 9a). We confirmed CD4-dependent vascular permeability to the DRG and the spinal cord using intravenous injection of 70 kDa fluorescein isothiocyanate (FITC)-dextran, which has a similar size to IgG (Extended Data Fig. 9b). Collectively, our results support the notion that CD4 T cells enable antibody delivery to the sites of infection by secreting IFN- γ and enhancing microvascular permeability. This mechanism of antibody delivery is crucial for host immune protection, as depletion of CD4 T cells, inhibition of CD4 T-cell migration into the neuronal tissues or neutralization of IFN- γ renders immune mice susceptible to infection.

To determine whether our findings extend beyond HSV-2, we examined antibody access to the neuronal tissue following a different neurotropic virus, vesicular stomatitis virus (VSV), a negative sense RNA virus of the Rhabdoviridae family. Upon intranasal inoculation, VSV infects olfactory sensory neurons in the nasal mucosa and enters the CNS through the olfactory bulb¹⁸. In contrast, intravenous infection with VSV is well tolerated, and generates robust T- and B-cell responses (Extended Data Fig. 10)¹⁹. To determine whether antibody access to the brain requires memory CD4 T cells, we immunized mice with VSV intravenously. Five weeks later, immunized mice were challenged with VSV intranasally. Entry of VSV-specific antibodies was monitored in the brain 6 days after intranasal challenge. Consistent with the data obtained from HSV-2 infection, we observed a striking dependence on CD4 T cells of antibody access to the brain (Extended Data Fig. 10b). Further, anti- α 4 antibody treatment of mice immediately before intranasal VSV challenge also diminished antibody access to the brain,

without impacting VSV-specific antibodies in circulation (Extended Data Fig. 10c). Furthermore, we observed that vascular permeability to the brain was dependent on $\alpha 4$ integrin, as antibody blockade of $\alpha 4$ integrin resulted in diminished albumin leakage to the brain (Extended Data Fig. 10d). Taken together, these results indicate that the requirement for $\alpha 4$ -integrin and memory CD4 T cells for antibody access applies to two distinct neurotropic viruses, HSV-2 and VSV, and suggest a general mechanism of antibody access to the immunoprivileged tissues protected by the blood–nerve barriers.

We have demonstrated a role of CD4 T cells in controlling antibody access to neuronal tissues through local migration and secretion of IFN- γ . Circulating CD4 memory T cells effectively target antibody delivery to the sites of infection through their secretion of IFN- γ , presumably upon recognition of cognate antigenic peptides presented by local antigen-presenting cells²⁰. These results indicate the requirement for CD4 T-cell help at the effector phase of the antibody response, and add to the growing appreciation of CD4 T cells in paving the way to other effector cell types such as CD8 T cells^{21–23}. We believe that the requirement for CD4 T cells for antibody access in neuronal tissue reflects an additional layer of control imposed by the immunoprivileged sites. In accessible tissues, inflammatory leukocytes can migrate and, in response to PAMPs, secrete cytokines such as TNF- α that are sufficient to trigger vascular permeability independently of CD4 T cells. However, after neurotropic viral infections, the infected neurons are expected to be poor at producing inflammatory cytokines that remodel vascular tight junctions. At the same time, recruitment of innate leukocytes is blocked by shutdown of specific chemokines in the ganglia of HSV-1-infected mice²⁴. Curiously, expression of T-cell-trophic chemokines CXCL9 and CXCL10 was preserved in the DRG of infected mice²⁴, suggesting that access by lymphocytes is permitted. Thus, in neuronal tissues, the entry of viral-specific CD4 T cells is crucial to provide cytokines that permit antibodies through the induction of vascular permeability.

On the other hand, aberrant entry and activation of CD4 T cells predispose immunoprivileged tissues for access to autoantibodies and tissue damage^{25,26}. Thus, this mode of targeted release of circulating antibodies not only provides a rapid and efficient mechanism of pathogen control, but may also restrict antibody release to irrelevant sites to limit immunopathology. Our results implicate that antibody-based vaccines or treatment against neurotropic viruses would benefit from generating robust circulating CD4 T-cell memory responses. Conversely, treatment of autoantibody-mediated neuropathies including chronic inflammatory demyelinating polyneuropathy and Guillain–Barré syndrome might benefit from preventing the accessibility of autoantibodies to target neurons enabled by CD4 T cells.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 14 November 2015; accepted 5 April 2016.

Published online 18 May 2016.

1. Hawkins, B. T. & Davis, T. P. The blood–brain barrier/neurovascular unit in health and disease. *Pharmacol. Rev.* **57**, 173–185 (2005).
2. Weerasuriya, A. & Mizisin, A. P. The blood–nerve barrier: structure and functional significance. *Methods Mol. Biol.* **686**, 149–173 (2011).
3. Koelle, D. M. & Corey, L. Herpes simplex: insights on pathogenesis and possible vaccines. *Annu. Rev. Med.* **59**, 381–395 (2008).
4. Knipe, D. M. & Cliffe, A. Chromatin control of herpes simplex virus lytic and latent infection. *Nature Rev. Microbiol.* **6**, 211–221 (2008).

5. Parr, M. B. *et al.* A mouse model for studies of mucosal immunity to vaginal infection by herpes simplex virus type 2. *Lab. Invest.* **70**, 369–380 (1994).
6. Iijima, N. & Iwasaki, A. T cell memory. A local macrophage chemokine network sustains protective tissue-resident memory CD4 T cells. *Science* **346**, 93–98 (2014).
7. Milligan, G. N., Bernstein, D. I. & Bourne, N. T lymphocytes are required for protection of the vaginal mucosae and sensory ganglia of immune mice against reinfection with herpes simplex virus type 2. *J. Immunol.* **160**, 6093–6100 (1998).
8. Parr, M. B. & Parr, E. L. Immunity to vaginal herpes simplex virus-2 infection in B-cell knockout mice. *Immunology* **101**, 126–131 (2000).
9. Sato, A. *et al.* Vaginal memory T cells induced by intranasal vaccination are critical for protective T cell recruitment and prevention of genital HSV-2 disease. *J. Virol.* **88**, 13699–13708 (2014).
10. Jones, C. A., Taylor, T. J. & Knipe, D. M. Biological properties of herpes simplex virus 2 replication-defective mutant strains in a murine nasal infection model. *Virology* **278**, 137–150 (2000).
11. Roopenian, D. C. & Akilesh, S. FcRn: the neonatal Fc receptor comes of age. *Nature Rev. Immunol.* **7**, 715–725 (2007).
12. McDermott, M. R., Brais, L. J. & Eveleigh, M. J. Mucosal and systemic antiviral antibodies in mice inoculated intravaginally with herpes simplex virus type 2. *J. Gen. Virol.* **71**, 1497–1504 (1990).
13. Morrison, L. A., Zhu, L. & Thebeau, L. G. Vaccine-induced serum immunoglobulin contributes to protection from herpes simplex virus type 2 genital infection in the presence of immune T cells. *J. Virol.* **75**, 1195–1204 (2001).
14. Ohashi, M., Bertke, A. S., Patel, A. & Krause, P. R. Spread of herpes simplex virus to the spinal cord is independent of spread to dorsal root ganglia. *J. Virol.* **85**, 3030–3032 (2011).
15. Man, S., Ubogu, E. E. & Ransohoff, R. M. Inflammatory cell migration into the central nervous system: a few new twists on an old tale. *Brain Pathol.* **17**, 243–250 (2007).
16. Anderson, K. G. *et al.* Intravascular staining for discrimination of vascular and tissue leukocytes. *Nature Protocols* **9**, 209–222 (2014).
17. Capaldo, C. T. *et al.* Proinflammatory cytokine-induced tight junction remodeling through dynamic self-assembly of claudins. *Mol. Biol. Cell* **25**, 2710–2719 (2014).
18. Reiss, C. S., Plakhov, I. V. & Komatsu, T. Viral replication in olfactory receptor neurons and entry into the olfactory bulb and brain. *Ann. NY Acad. Sci.* **855**, 751–761 (1998).
19. Thomsen, A. R. *et al.* Cooperation of B cells and T cells is required for survival of mice infected with vesicular stomatitis virus. *Int. Immunol.* **9**, 1757–1766 (1997).
20. Iijima, N. *et al.* Dendritic cells and B cells maximize mucosal Th1 memory response to herpes simplex virus. *J. Exp. Med.* **205**, 3041–3052 (2008).
21. Laidlaw, B. J. *et al.* CD4⁺ T cell help guides formation of CD103⁺ lung-resident memory CD8⁺ T cells during influenza viral infection. *Immunity* **41**, 633–645 (2014).
22. Nakanishi, Y., Lu, B., Gerard, C. & Iwasaki, A. CD8⁺ T lymphocyte mobilization to virus-infected tissue requires CD4⁺ T-cell help. *Nature* **462**, 510–513 (2009).
23. Reboldi, A. *et al.* C-C chemokine receptor 6-regulated entry of TH-17 cells into the CNS through the choroid plexus is required for the initiation of EAE. *Nature Immunol.* **10**, 514–523 (2009).
24. Stock, A. T., Smith, J. M. & Carbone, F. R. Type I IFN suppresses Cxcr2 driven neutrophil recruitment into the sensory ganglia during viral infection. *J. Exp. Med.* **211**, 751–759 (2014).
25. Westland, K. W. *et al.* Activated non-neural specific T cells open the blood–brain barrier to circulating antibodies. *Brain* **122**, 1283–1291 (1999).
26. Pollard, J. D. *et al.* Activated T cells of nonneural specificity open the blood–nerve barrier to circulating antibody. *Ann. Neurol.* **37**, 467–475 (1995).

Acknowledgements We thank H. Dong, S. L. Fink and K. Hashimoto-Torii for animal care support and technical help. We thank R. Medzhitov for discussions. This study was supported by awards from National Institutes of Health grants AI054359, AI062428, AI064705 (to A.I.). A.I. is an investigator of the Howard Hughes Medical Institute.

Author Contributions N.I. and A.I. planned the project, designed experiments, analysed and interpreted data and wrote the manuscript. N.I. performed experiments.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.I. (akiko.iwasaki@yale.edu).

METHODS

Mice. Six- to eight-week-old female C57BL/6 (CD45.2⁺) and congenic C57BL/6 B6.SJL-PtprcaPep3b/BoyJ (B6.Ly5.1) (CD45.1⁺) mice, B6.129S2-Igh^{tm1Cgn}/J (μ MT) mice, anti-HEL B-cell receptor (BCR)-transgenic C57BL/6-TgN (IghelMD4) (HELtG) mice, CBy.PL(B6)-Thy1^a/ScrJ (Thy1.1⁺ BALB/c) mice and B6.129X1-Fcgrt^{tm1Dcr}/DcrJ (FcRn^{-/-}) mice were purchased from the National Cancer Institute and Jackson Laboratory. J_HD mice (B-cell deficient on BALB/c background) were obtained from Taconic Animal Models. All procedures used in this study complied with federal guidelines and institutional policies by the Yale School of Medicine Animal Care and Use Committee.

Viruses. HSV-2 strains 186syn⁻ TK⁻ and 186syn⁺ were gifts from D. Knipe. These viruses were propagated and titred on Vero cells (ATCC CCL-81) as previously described²⁰. Influenza virus A/Puerto Rico/3/334 (A/PR8: H1N1) and WT/VSV were propagated as previously described^{20,27}.

Virus infection. Six- to eight-week-old female mice injected subcutaneously with Depo Provera (Pharmacia Upjohn, 2 mg per mouse) were immunized intravaginally, intraperitoneally or intranasally with 10⁵ p.f.u. of HSV-2 (186syn⁻ TK⁻) as previously described⁶. For secondary challenge, immunized mice were challenged vaginally with 10⁴ p.f.u. of WT HSV-2 (186syn⁺) (100% lethal dose for naive mice). In the case of BALB/c and J_HD mice, these mice were immunized with 5 × 10⁴ to 10⁵ p.f.u. of HSV-2. For secondary challenge, immunized mice were challenged with 10⁵ p.f.u. of WT HSV-2 (100% lethal dose for naive mice). The severity of disease was scored as follows: 0, no sign; 1, slight genital erythema and oedema; 2, moderate genital inflammation; 3, purulent genital lesions; 4, hind-limb paralysis; 5, pre-moribund²⁰. Owing to humane concerns, the animals were euthanized before reaching moribund state. To measure virus titre in peripheral tissues, vaginal tissues, DRG and spinal cord were harvested in ABC buffer (0.5 mM MgCl₂·6H₂O, 0.9 mM CaCl₂·2H₂O, 1% glucose, 5% HI FBS and penicillin-streptomycin) including 1% amphotericin-B (Sigma). Thereafter, these tissues were homogenized by lysing matrix D (MP Biomedicals), followed by clarifying by centrifugation. Viral titres were obtained by titration of tissue samples on a Vero cell monolayer. Protein concentration in tissue homogenates was measured by a DC protein assay kit (Bio-Rad Laboratories). C57BL/6 mice were immunized intravenously with WT/VSV (2 × 10⁶ p.f.u. per mouse) or intranasally with influenza A/PR8 (10 p.f.u. per mouse). For secondary challenge, VSV-immunized mice were re-infected intranasally with WT/VSV (1 × 10⁷ p.f.u. per mouse).

Antibodies. Anti-CD90.2 (30-H12), anti-CD90.1 (OX-7), anti-CD45.2 (104), anti-CD45.1 (A20), anti-CD4 (GK1.5, RM4-5 and RM4-4), anti-CD19 (6D5), anti-CD45R/B220 (RA3-6B2), anti-MHC class II (I-A/I-E, M5/114.15.2), anti-CD69 (H1.2F3), anti-CD44 (IM7), anti-CD49d (R1-2), anti-NKp46 (29A1.4) and anti-IFN- γ (XMG1.2 and R4-6A2) were purchased from e-Bioscience or Biolegend.

Isolation of leukocytes from peripheral tissues. The genital tracts of vaginal tissues treated with Depo-Provera were dissected from the urethra and cervix. Before collection of neuronal tissues, mice were perfused extensively using transcardiac perfusion and perfusion through inferior vena cava and great saphenous vein with more than 30 ml of PBS. The DRG and the adjacent region of the spinal cord were harvested in PBS for flow cytometry or ABC buffer for tissue homogenization. The tissues in PBS were then incubated with 0.5 mg ml⁻¹ Dispase II (Roche) for 15 min at 37°C. Thereafter, vaginal tissues were digested with 1 mg ml⁻¹ collagenase D (Roche) and 30 μ g ml⁻¹ DNase I (Sigma-Aldrich) at 37°C for 25 min. The resulting cells were filtered through a 70- μ m filter^{28,29}.

Flow cytometry. Preparation of single-cell suspensions from spleen, draining lymph nodes (inguinal lymph node and iliac lymph nodes), vagina and neuronal tissues were described previously. Multiparameter analyses were performed on an LSR II flow cytometer (Becton Dickinson) and analysed using FlowJo software (Tree Star). To detect HSV-2-specific CD4⁺ T cells or VSV-specific CD4⁺ T cells (CD45.1⁺ or CD45.2⁺), single-cell suspensions from vaginal tissues of TK⁻ HSV-2-immunized mice or VSV immunized mice were stimulated in the presence of 5 μ g ml⁻¹ Brefeldin A with naive splenocytes (CD45.1⁺CD45.2⁺) loaded with heat-inactivated HSV-2 antigen, heat-inactivated WT VSV or heat-inactivated influenza virus A/PR8 for around 12 h (ref. 6). To detect HSV-2-specific CD4⁺ T cells in BALB/c and J_HD mice, single-cell suspensions (CD90.2⁺) from vaginal tissues of TK⁻ HSV-2-immunized mice were stimulated with naive splenocytes (CD90.1⁺) loaded with heat-inactivated HSV-2 antigen.

In vivo treatment with neutralizing/depleting antibodies. C57BL/6 mice or BALB/c mice were immunized with TK⁻ HSV-2 virus. Five to eight weeks later, these mice were injected intravenously (tail vein) with 300 μ g of anti-CD4 (GK1.5; BioXCell) or anti-IFN- γ (XMG1.2; BioXCell) antibody at days -4, -1, 2 and 4 after HSV-2 challenge. *In vivo* depletion for CD4 was confirmed by fluorescence-activated cell sorting analysis of the cell suspension from spleen. For the

neutralization of α 4-integrin, purified anti-mouse α 4 integrin/CD49d (PS/2; SouthernBiotech) was given a tail vein injection of 300 μ g antibody at days 2 and 4 after challenge.

Parabiosis. Parabiosis was performed as previously described with slight modifications⁶. Naive or immunized C57BL/6 mice, HELtG and μ MT mice were anaesthetized with a mixture of ketamine/xylazine (100 mg/kg and 10 mg/kg body weight respectively). After shaving the corresponding lateral aspects of each mouse, matching skin incisions were made from behind the ear to hip and sutured together with Chromic Gut (4-0, Henry Schein) absorbable suture, then these areas were clipped with 7-mm stainless-steel wound clips (Roboz).

Measurement of virus-specific Ig and total Ig in serum and tissue homogenates. Ninety-six-well EIA/RIA plates were filled with 100 μ l of heat-inactivated purified HSV-2 (10⁴–10⁵ p.f.u. equivalent per 100 μ l) or heat-inactivated purified VSV (5 × 10⁵ p.f.u. equivalent per 100 μ l) for virus-specific Ig measurement or goat anti-mouse Ig (1:1,000; SouthernBiotech, 1010-01) for total Ig measurement in carbonate buffer (pH 9.5) and then incubated overnight at 4°C. On the following day, these plates were washed with PBS-Tween 20 and blocked for 2 h with 5% FBS in PBS. Tissue samples and serum samples in ABC buffer were then plated in the wells and incubated for at least 4 h at ambient temperature (20–25°C). After washing in PBS-Tween 20, HRP-conjugated anti-mouse IgG1, IgG3, IgM, IgA, IgG2a, IgG2b or IgG2c (SouthernBiotech) was added to the wells for 1 h, followed by washing and adding TMB solution (eBioscience). Reactions were stopped with 1 N H₂SO₄ and absorbance was measured at 450 nm. The sample antibody titres were defined by using Ig standard (C57BL/6 Mouse Immunoglobulin Panel; SouthernBiotech) or mouse IgG2a (HOPC-1; SouthernBiotech).

Albumin ELISA. Using tissue homogenates (DRG and spinal cord) prepared after extensive perfusion, albumin ELISA (Genway) was performed according to instruction.

Immunofluorescence staining. Frozen sections 8 μ m in thickness were cut, fixed and left to dry at ambient temperature. These tissues were stained with the antibodies (anti-CD4 (H129.19), anti-MHC class II (M5/114.15.2) anti-VCAM-1 (429/MVCAM.A), anti-CD31 (390 and MEC13.3), anti-Ly6G (1A8), anti-CD11b (M1/70) and anti-mouse albumin (Goat pAb/Bethyl Laboratories)) as previously described⁶. These slides were washed and incubated with DAPI and mounted with Fluoromount-G (SouthernBiotech). They were analysed by fluorescence microscopy (BX51; Olympus).

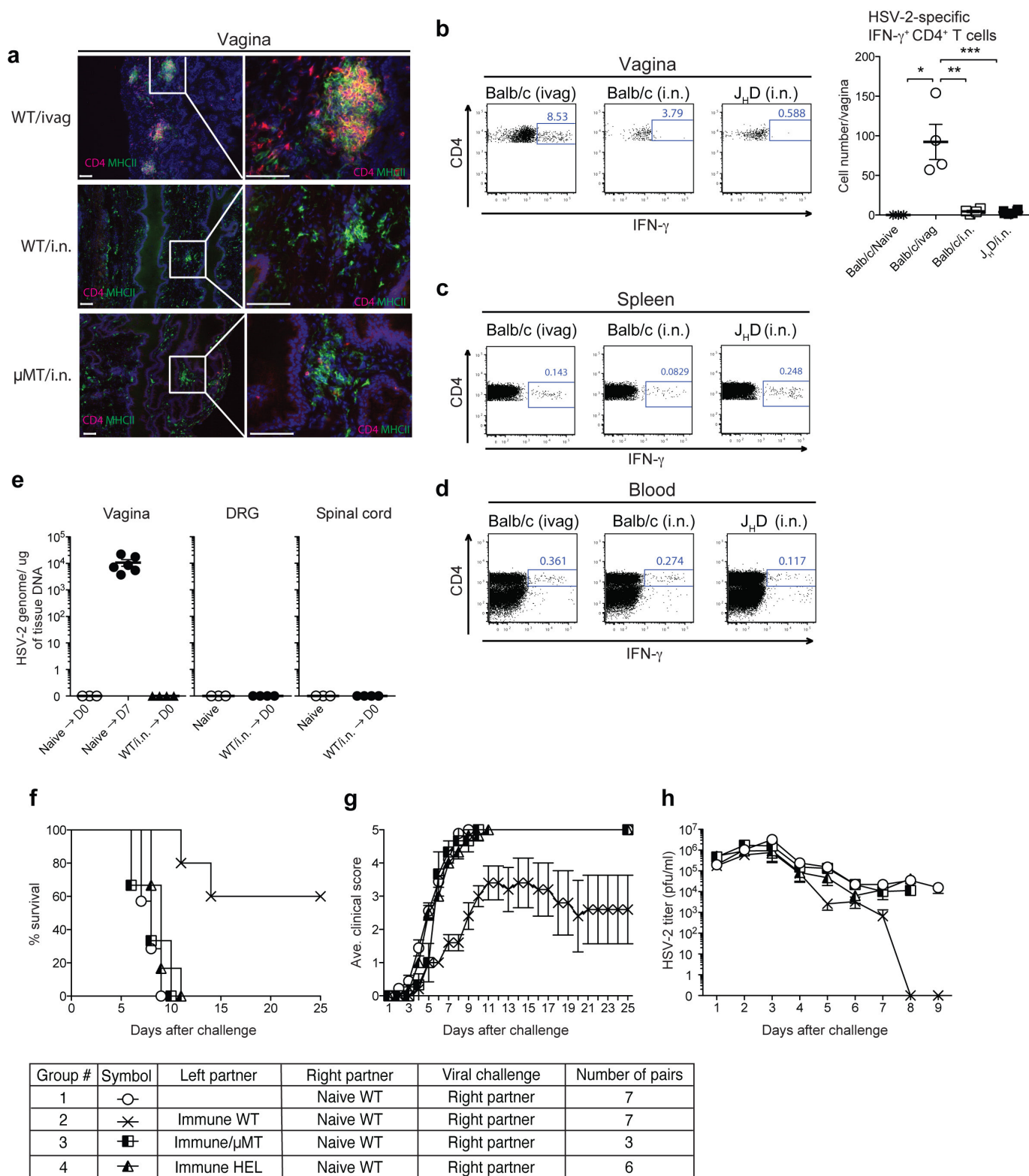
Vascular permeability assays. Spinal column was harvested from intranasal TK⁻ HSV-2-immunized mice 45 min after tail vein injection with 200 μ l of 5 mg ml⁻¹ Oregon Green 488-conjugated dextran (70 kDa, D7173, Thermo Fisher Scientific) in PBS. Spine was then fixed with 4% paraformaldehyde in PBS overnight, and frozen sections cut (8 μ m in thickness) for immunohistochemical analysis³⁰.

DNA isolation from tissues. C57BL/6 mice were immunized intranasally with TK⁻ HSV-2. Six weeks later, vaginal tissues, DRG and spinal cord of these mice were lysed in 10 mg ml⁻¹ Proteinase K (Roche) to isolate DNA at 55°C overnight. After removing these tubes, phenol equilibrated with Tris pH 8.0 was added. Thereafter, upper aqueous phase was added to phenol/chloroform (1:1). The upper aqueous phase was re-suspended with sodium acetate, pH 6.0, and 100% ethanol at room temperature. After shaking and centrifuging, the concentration of isolated DNA pellet was measured. The level of HSV-2 genomic DNA in peripheral tissues on the basis of HSV-2 gD (forward primer: AGCGAGGATAACCTGGGATT; reverse primer: GGGATAAAGCGGGGTACAT) was analysed by quantitative PCR using purified viral DNA genome as standard.

Statistical analysis. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Survival curves were analysed using a log-rank test. For other data, normally distributed continuous variable comparisons used a two-tailed unpaired Student's *t*-test or paired Student's *t*-test with Prism software. To compare two non-parametric data sets, a Mann-Whitney *U*-test was used.

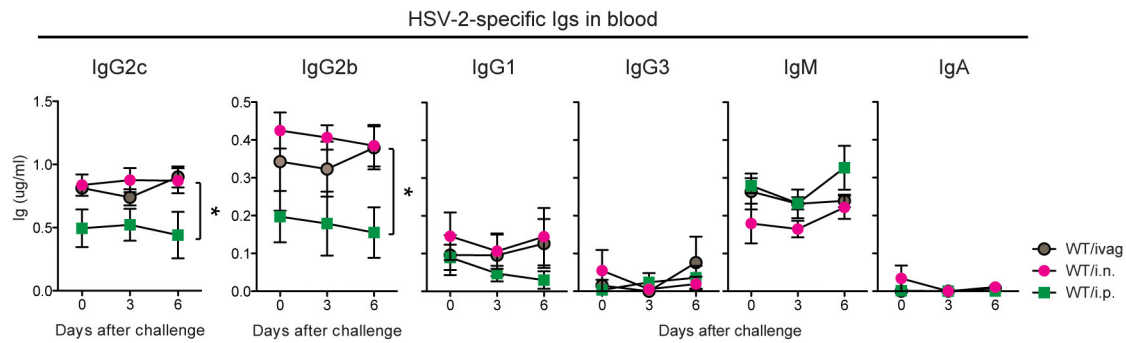
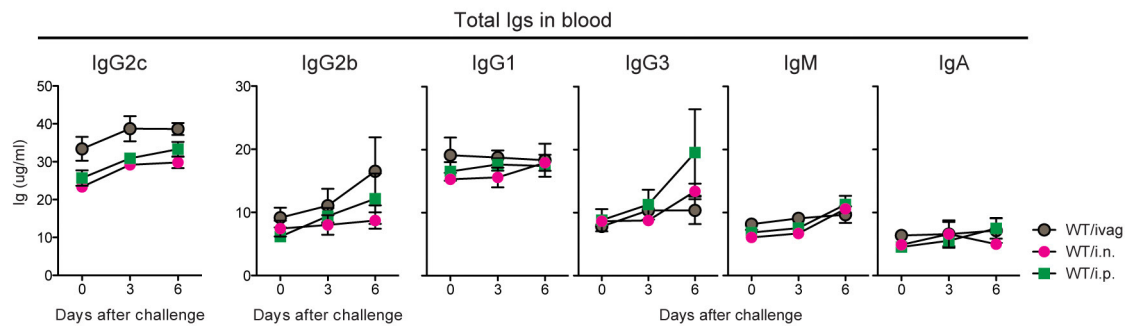
27. Sasai, M., Linehan, M. M. & Iwasaki, A. Bifurcation of Toll-like receptor 9 signaling by adaptor protein 3. *Science* **329**, 1530–1534 (2010).
28. Iijima, N., Mattei, L. M. & Iwasaki, A. Recruited inflammatory monocytes stimulate antiviral Th1 immunity in infected tissue. *Proc. Natl Acad. Sci. USA* **108**, 284–289 (2011).
29. Johnson, A. J., Chu, C. F. & Milligan, G. N. Effector CD4⁺ T-cell involvement in clearance of infectious herpes simplex virus type 1 from sensory ganglia and spinal cords. *J. Virol.* **82**, 9678–9688 (2008).
30. Knowland, D. et al. Stepwise recruitment of transcellular and paracellular pathways underlies blood-brain barrier breakdown in stroke. *Neuron* **82**, 603–617 (2014).



Extended Data Figure 1 | In the absence of T_{RM}, B cells are required for the protection of the host against genital HSV-2 challenge. **a**, C57BL/6 mice and μMT mice were immunized intravaginally or intranasally with TK⁻ HSV-2. Five weeks later, vaginal tissue sections were stained for CD4⁺ cells (red) and MHC class II⁺ cells (green). Blue labelling depicts nuclear staining with DAPI (blue). Images were captured using a $\times 10$ or $\times 40$ objective lens. Scale bars, 100 μ m. Data are representative of three similar experiments. **b–d**, BALB/c mice and J_HD mice were immunized with TK⁻ HSV-2 (10^5 p.f.u.) intranasally or intravaginally. Six weeks later, the number of total CD4⁺ T cells and HSV-2-specific IFN- γ ⁺ CD4⁺ T cells in the vagina (**b**), spleen (**c**) and peripheral blood (**d**) were analysed by flow cytometry. Percentages and total number of IFN- γ ⁺ cells among

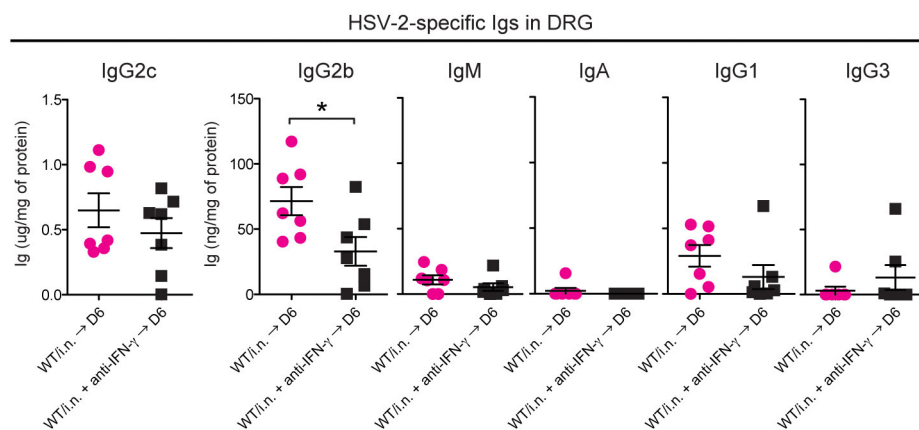
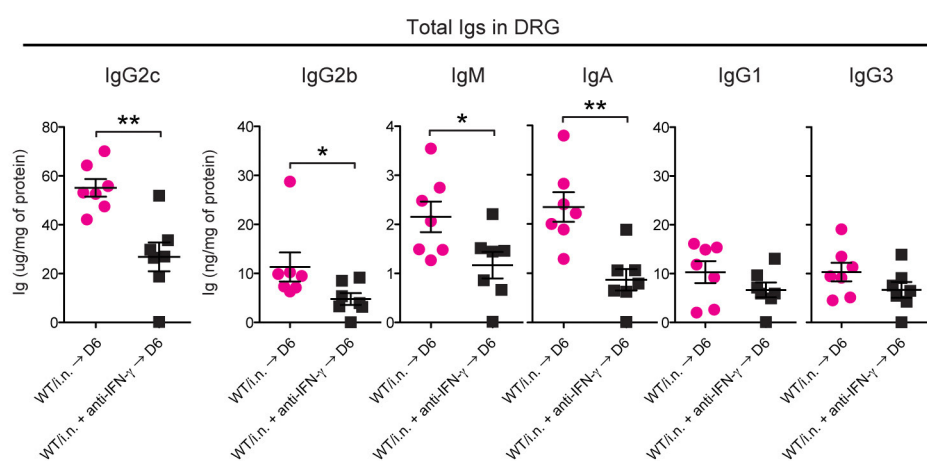
CD4⁺CD90.2⁺ cells are shown. Data are mean \pm s.e.m. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$ (two-tailed unpaired Student's *t*-test).

e, C57BL/6 mice were immunized intravaginally (naive \rightarrow D7) or intranasally (WT/i.n. \rightarrow D0) with TK⁻ HSV-2 virus. At the indicated time points (D7: 7 days after immunization; WT/i.n. \rightarrow D0: 6 weeks after immunization), total viral genomic DNA in the vaginal tissues, DRG and spinal cord were measured by quantitative PCR. **f–h**, Intravaginally immunized C57BL/6 (WT), μMT and HEL-BCR Tg mice (left partner) were surgically joined with naive WT mice (right partner). Three weeks after parabiosis, the naive partner was challenged with a lethal dose of WT HSV-2 intravaginally. Mortality (**e**), clinical score (**f**) and virus titre in vaginal wash (**g**) following viral challenge are depicted.

a**b**

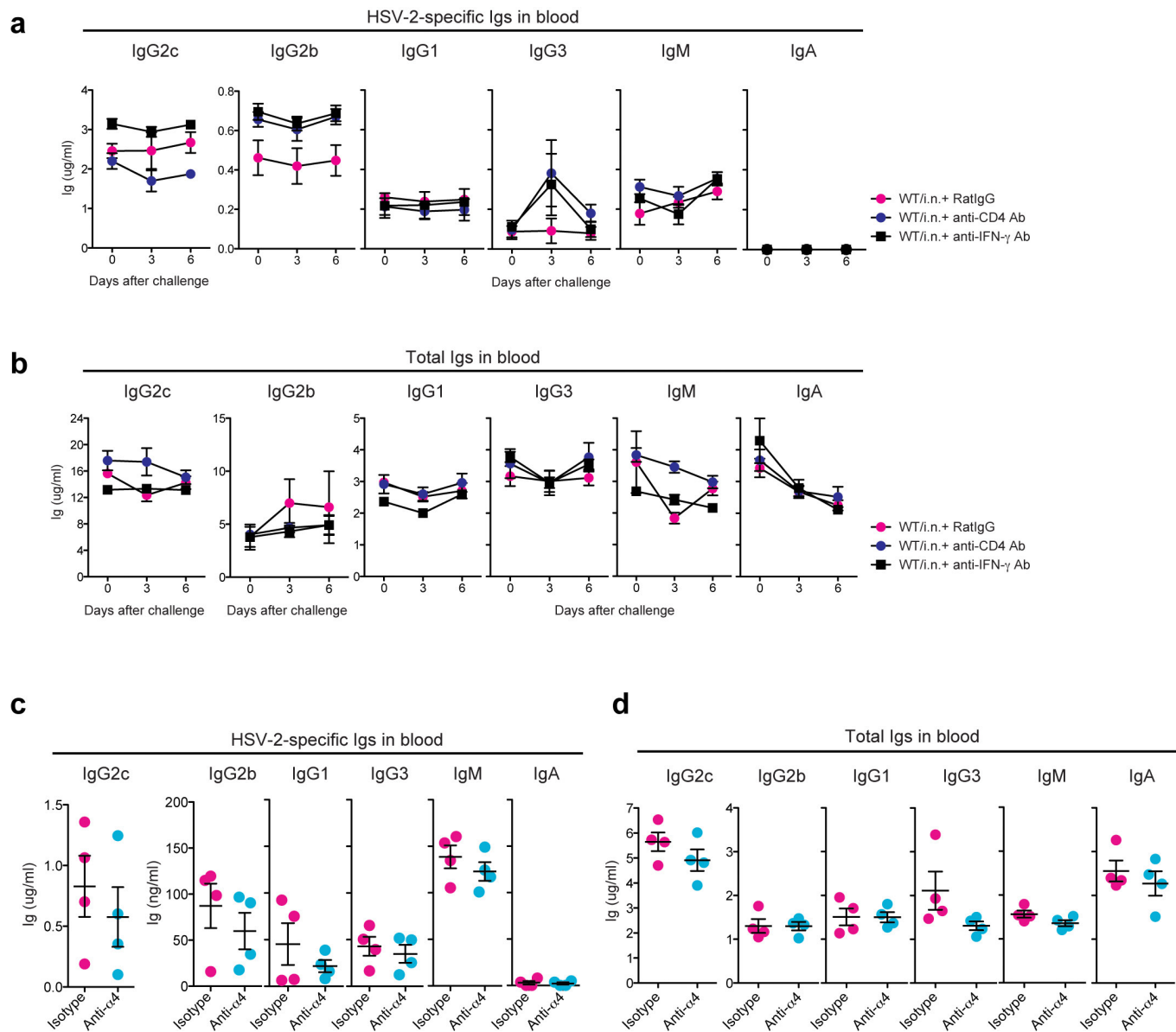
Extended Data Figure 2 | Mucosal TK⁻ HSV-2 immunization generates higher levels of virus-specific IgG2b and IgG2c compared with intraperitoneal immunization. WT mice were immunized with TK⁻ HSV-2 (10^5 p.f.u. per mouse) via intravaginal, intraperitoneal or intranasal

routes. Six weeks later, these mice were challenged with a lethal dose of WT HSV-2 intravaginally. At the indicated days after challenge, HSV-2-specific Ig (a) and total Ig (b) in serum were analysed by ELISA. Data are mean \pm s.e.m. * $P < 0.05$ (Mann-Whitney *U*-test).

a**b**

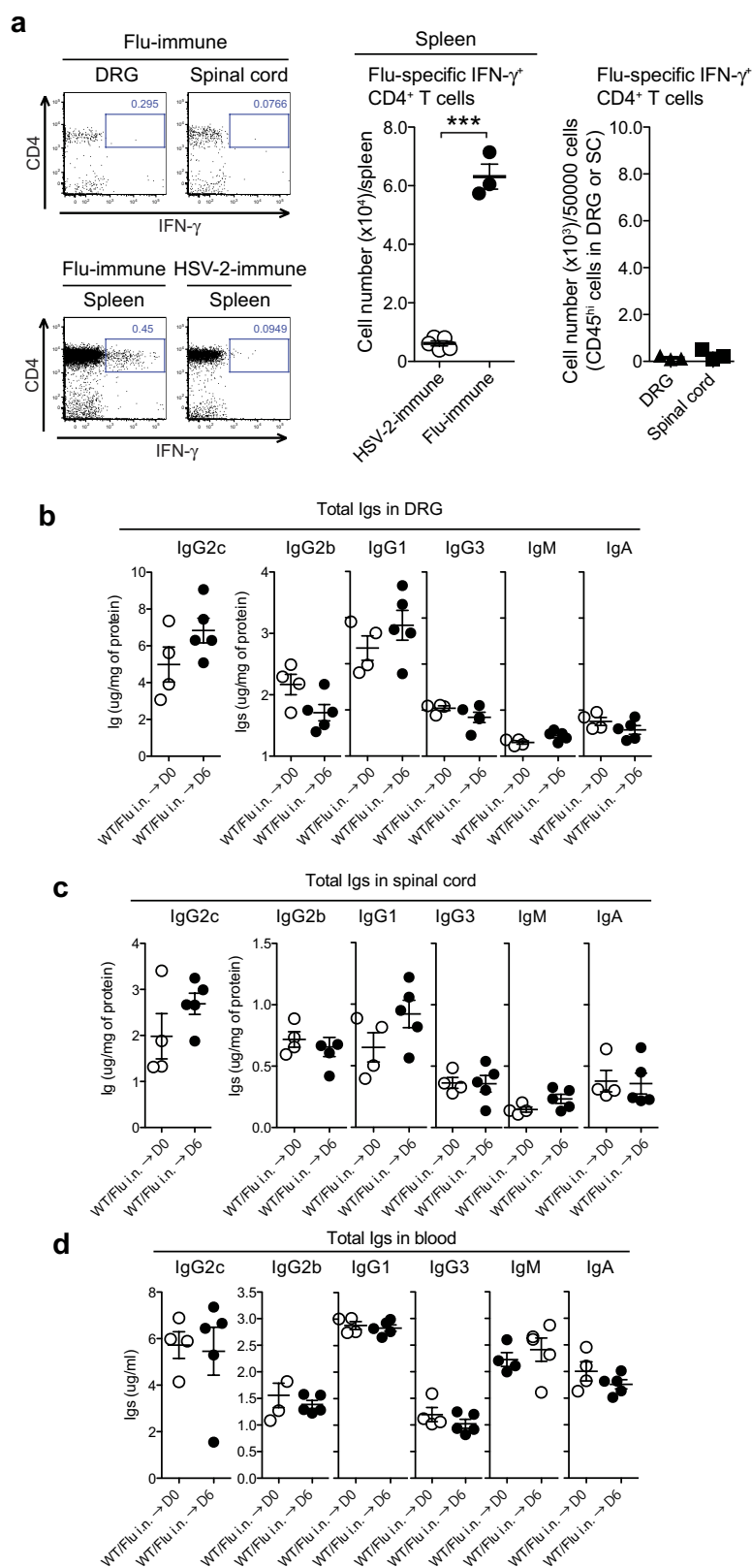
Extended Data Figure 3 | IFN- γ enhances antibody access to the DRG. WT mice immunized with TK⁻ HSV-2 (10^5 p.f.u. per mouse) intranasally 6 weeks earlier were challenged with a lethal dose of WT HSV-2 intravaginally. Six days after challenge, after extensive perfusion, HSV-2-specific (a) and total Ig (b) in DRG homogenates were analysed by

ELISA. Depletion of CD4 T cells or neutralization of IFN- γ was performed on days -4, and -1, 2 and 4 days after challenge by intravenous injection of anti-CD4 (GK1.5) or anti-IFN- γ (XMG1.2), respectively. Data are mean \pm s.e.m. * P < 0.05; ** P < 0.001 (two-tailed unpaired Student's t -test).



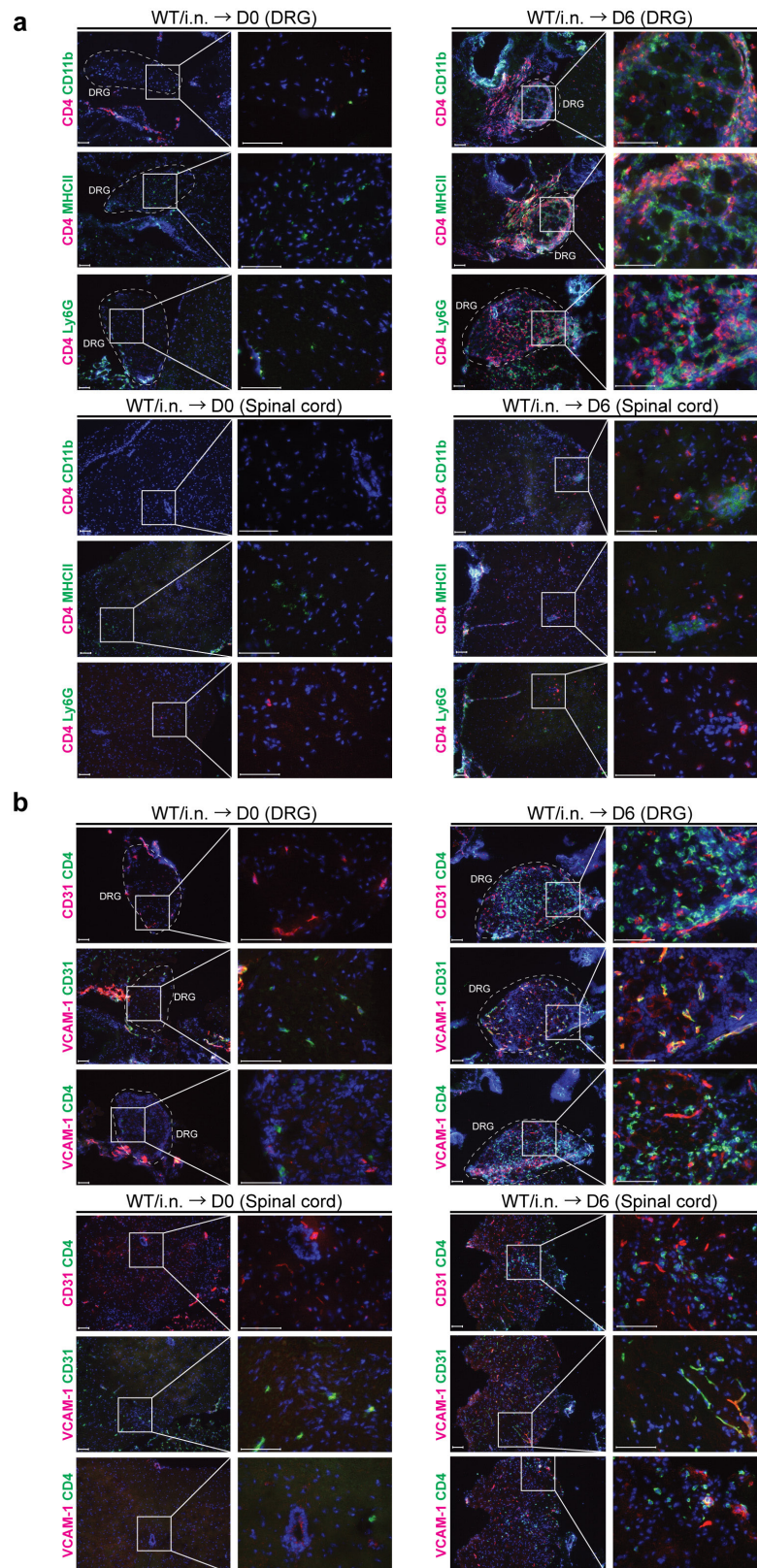
Extended Data Figure 4 | Neutralization of IFN- γ , α 4-integrin or depletion of CD4 T cells has no impact on circulating immunoglobulin levels. **a, b**, WT mice immunized intranasally with TK⁻ HSV-2 6–8 weeks earlier were challenged with a lethal dose of WT HSV-2. Depletion of CD4 T cells or neutralization of IFN- γ was performed on days –4, and –1, 2 and 4 days after challenge by intravenous injection of anti-CD4 (GK1.5) or anti-IFN- γ (XMG1.2), respectively. At time points indicated, HSV-2-specific Ig in the blood ($n = 4$) (**a**) and total Ig in the blood

($n = 4$) (**b**) were measured. **c, d**, WT mice immunized intranasally with TK⁻ HSV-2 6 weeks earlier were challenged with a lethal dose of WT HSV-2. Neutralization of α 4-integrin was performed on days 2 and 4 after challenge by intravenous injection of anti- α 4-integrin/CD49b antibody. Six days later, HSV-2-specific antibody (**c**) and total antibody (**d**) in the blood were measured. Data are representative of three similar experiments.



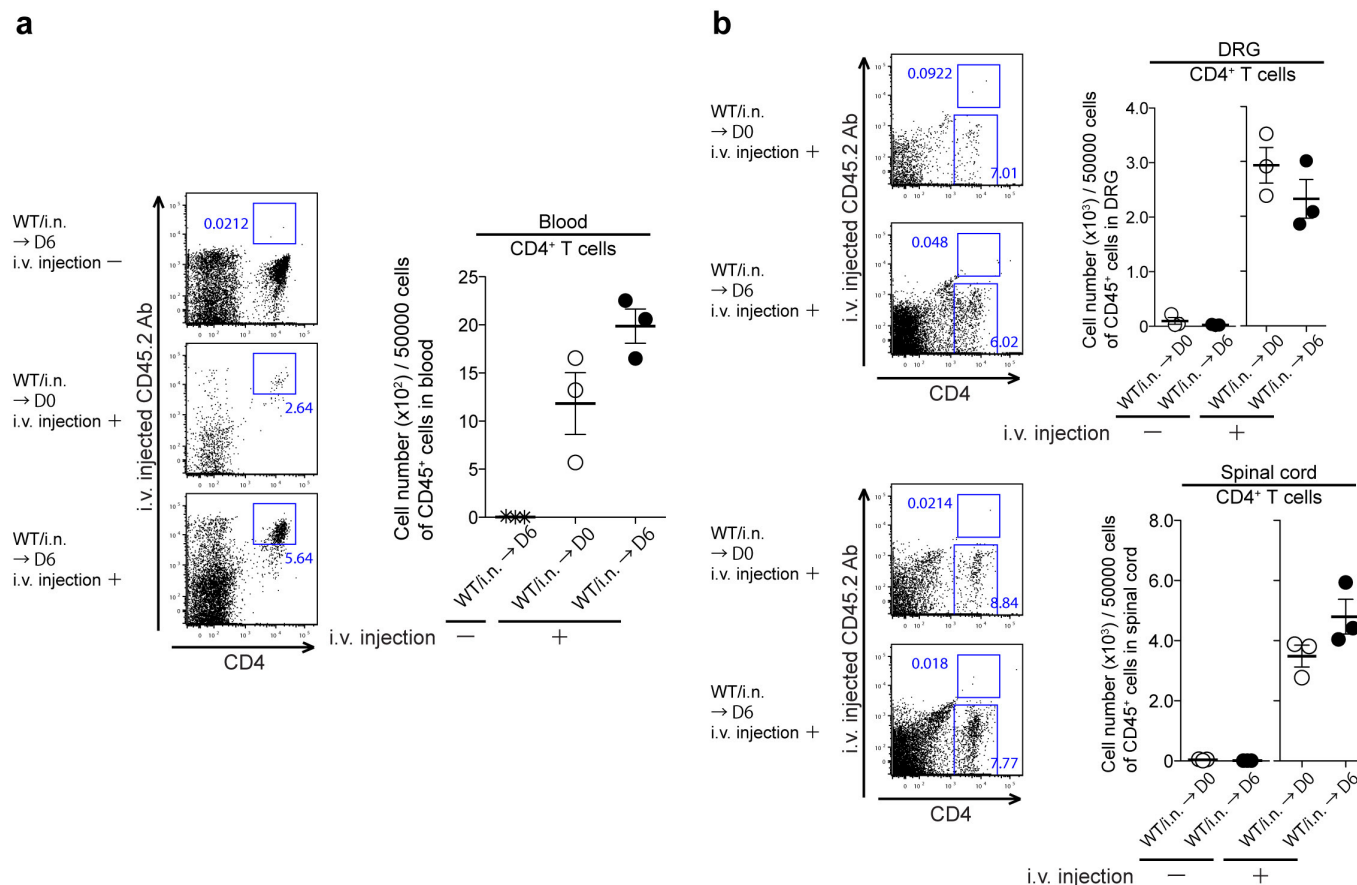
Extended Data Figure 5 | An irrelevant immunization fails to increase the levels of total antibodies in neuronal tissues. **a**, C57BL/6 mice were immunized with a sublethal dose of influenza A/PR8 virus (10 p.f.u. per mouse) intranasally. Three weeks later, Flu-specific IFN- γ ⁺ CD4⁺ T cells in spleen and neuronal tissues (DRG and spinal cord) (CD45.2⁺) following co-culture with HI-Flu/PR8 loaded splenocytes (CD45.1⁺) were analysed by flow cytometry. As a control, lymphocytes isolated from spleen of

TK⁻ HSV-2 intranasally immunized mice 6 weeks after vaccination were used for co-culture. (***) $P < 0.001$; two-tailed unpaired Student's *t*-test). **b–d**, C57BL/6 mice were immunized with a sublethal dose of influenza A/PR8 virus (10 p.f.u. per mouse). Four weeks later, these mice were challenged with a lethal dose of WT HSV-2 (10⁴ p.f.u. per mouse) intravaginally. Six days after challenge, total antibodies in lysate in DRG (**b**), spinal cord (**c**) and blood (**d**) were measured by ELISA.



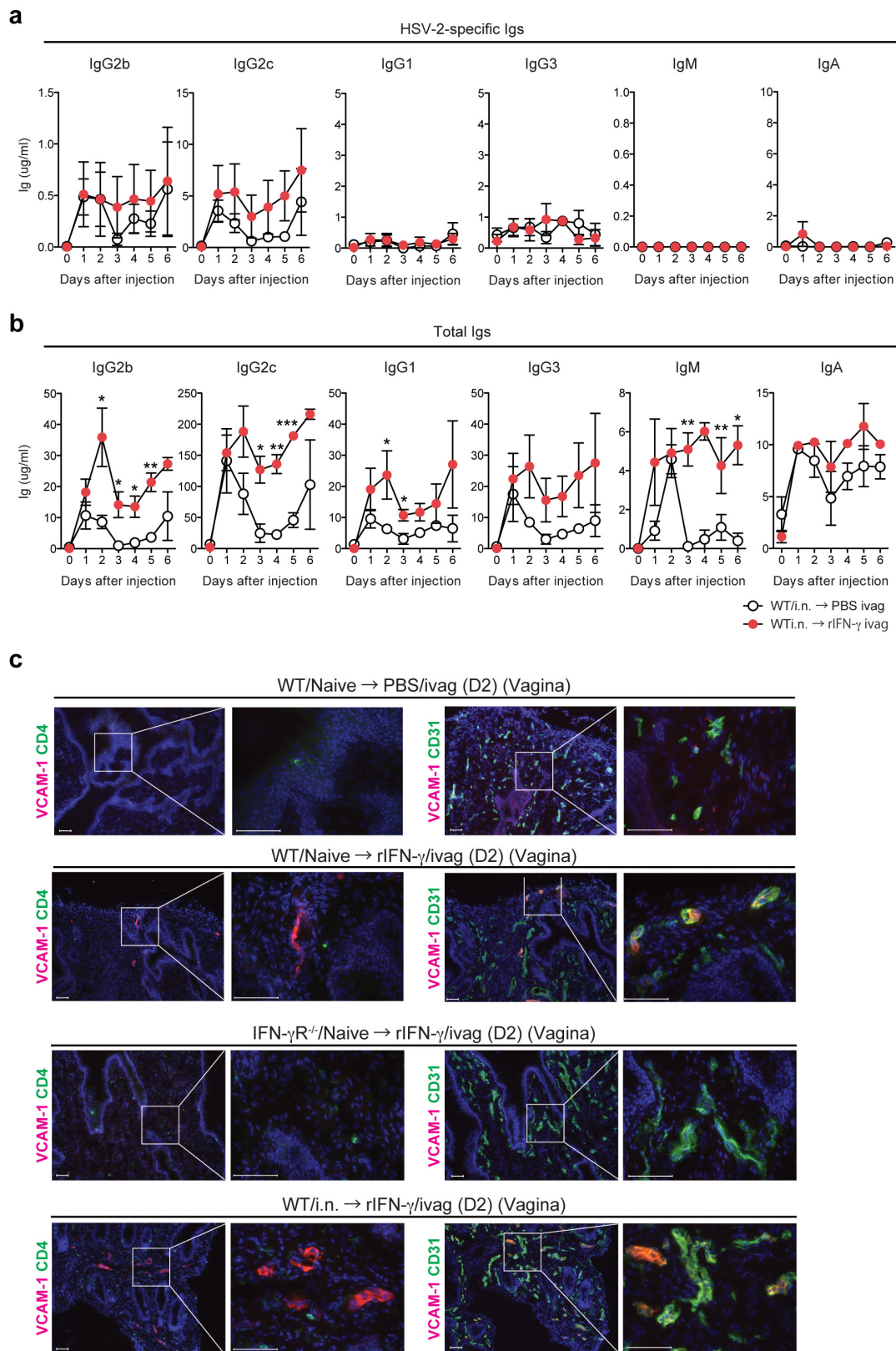
Extended Data Figure 6 | Most CD4 T cells recruited to the DRG and spinal cord of immunized mice are localized in the parenchyma of neuronal tissues. **a**, C57BL/6 mice were immunized intranasally with TK⁻ HSV-2. Six days after challenge of immunized mice 6 weeks prior, neuronal tissue sections (DRG and spinal cord) were stained for CD4⁺ cells and VCAM-1⁺ cells or CD31⁺ cells (red or green). Blue labelling depicts nuclear staining with DAPI (blue). Images were captured using a $\times 10$ or $\times 40$ objective lens. Scale bars, 100 μ m. **b**, C57BL/6 mice were

immunized intranasally with TK⁻ HSV-2. Six weeks later, mice were challenged with WT HSV-2 intravaginal and neuronal tissues were collected 6 days later. DRG and spinal cord were stained for CD4⁺ cells (red) and MHC class II⁺ cells, CD11b⁺ cells or Ly6G⁺ cells (green). Blue labelling depicts nuclear staining with DAPI (blue). Images were captured using a $\times 10$ or $\times 40$ objective lens. Scale bars, 100 μ m. Data are representative of at least three similar experiments.



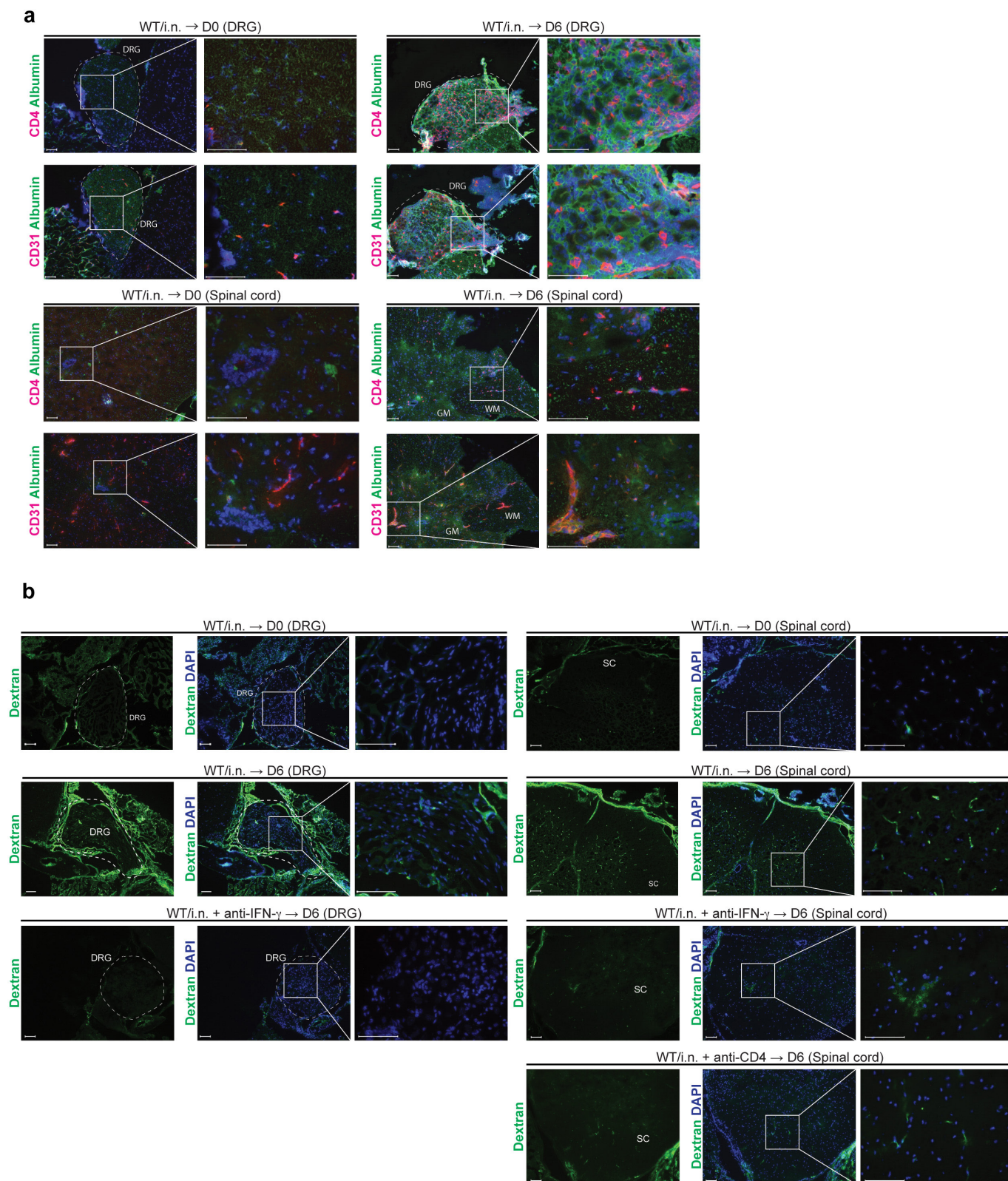
Extended Data Figure 7 | Intravascular staining reveals localization of CD4 T cells in the parenchyma of neuronal tissues. **a, b,** C57BL/6 mice immunized intranasally with TK⁻ HSV-2 6 weeks previously were challenged with lethal WT HSV-2. Six days after challenge, Alexa Fluor 700-conjugated anti-CD90.2 antibody (3 μ g per mouse) was injected

intravenously (tail vein) into immunized mice. Five minutes later, these mice were killed for fluorescence-activated cell sorting analysis of intravascular versus extravascular lymphocytes. Data are representative of at least two similar experiments.



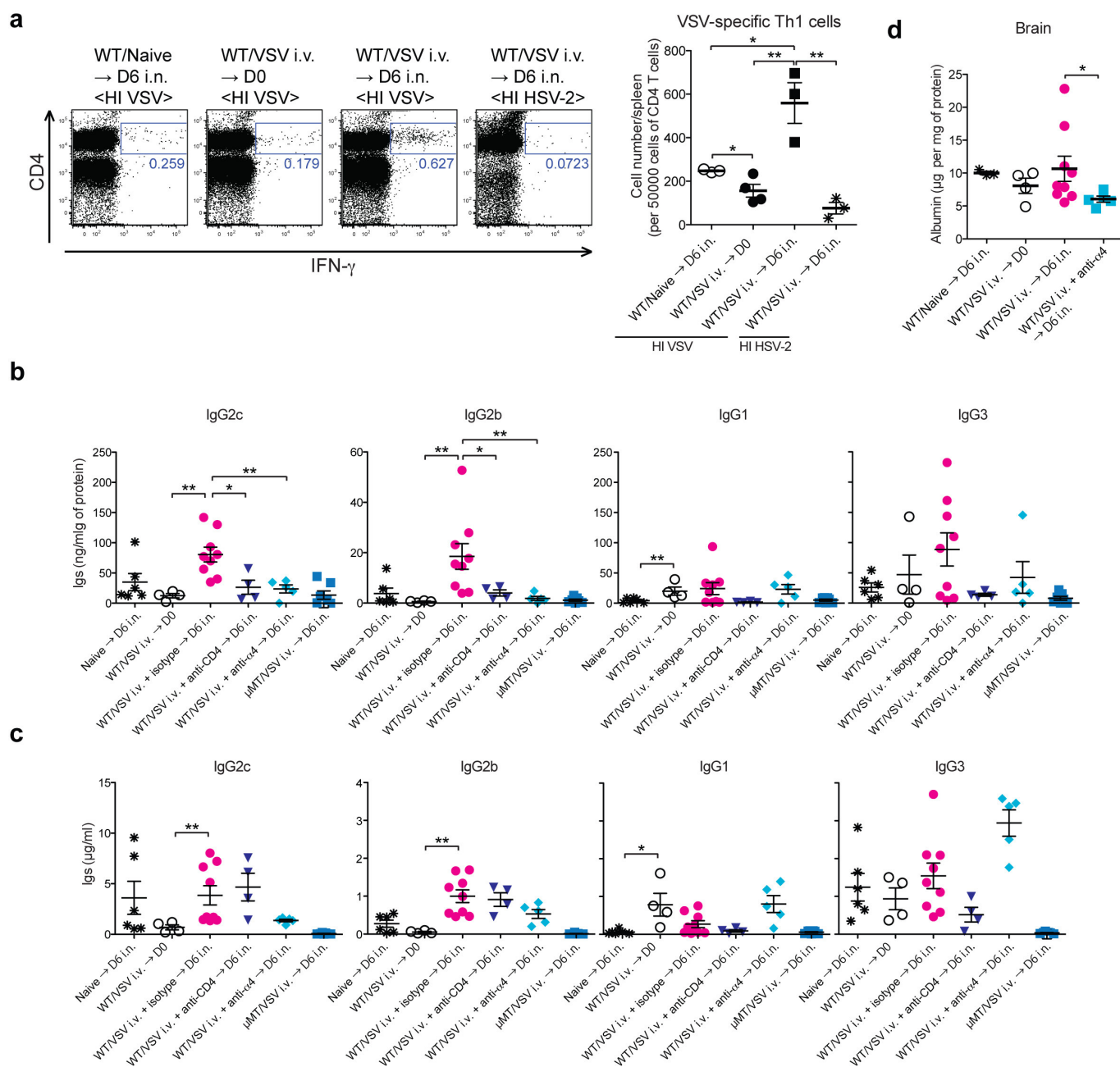
Extended Data Figure 8 | Recombinant IFN- γ is sufficient to increase epithelial and vascular permeability in vaginal tissues. **a**, WT mice immunized with TK⁻ HSV-2 (10^5 p.f.u.) intranasally 6 weeks earlier were injected intravaginally with recombinant mouse IFN- γ ($10\ \mu\text{g}$ per mouse) ($n = 3$) or PBS ($n = 3$). At the indicated time points, HSV-2-specific Ig (**a**) and total Ig (**b**) in vaginal wash were measured by ELISA. **c**, Two days after

rIFN- γ treatment, vaginal tissue sections were stained for VCAM-1⁺ cells (red) or CD4⁺ cells (green) and CD31⁺ cells (green). Blue labelling depicts nuclear staining with DAPI (blue). Images were captured using a $\times 10$ or $\times 40$ objective lens. Scale bars, $100\ \mu\text{m}$. Data are representative of at least three similar experiments.



Extended Data Figure 9 | Vascular permeability in DRG and spinal cord is augmented following WT HSV-2 challenge. **a**, C57BL/6 mice were immunized intranasally with TK⁻ HSV-2. Six days after challenge of mice immunized 6 weeks previously, neuronal tissue sections (DRG and spinal cord) were stained for CD4⁺ cells (red) and mouse albumin (green). Blue labelling depicts nuclear staining with DAPI (blue). **b**, C57BL/6 mice were immunized intranasally with TK⁻ HSV-2. Six weeks later, these mice

were challenged with lethal WT HSV-2. Six days after challenge, Oregon green 488-conjugated dextran (70 kDa) (5 mg ml⁻¹, 200 μ l per mouse) was injected intravenously into intranasally immunized mice. Forty-five minutes later, these mice were killed for immunohistochemical analysis. GM, grey matter; WM, white matter. Data are representative of three similar experiments.



Extended Data Figure 10 | Memory CD4⁺ T cells are required for the increase in antibody levels and vascular permeability in the brain following VSV immunization and challenge. **a**, C57BL/6 mice were immunized intravenously with WT VSV (2×10^6 p.f.u. per mouse). Five weeks later, these mice were challenged intranasally with WT VSV (1×10^7 p.f.u. per mouse). Six days after challenge, VSV-specific IFN- γ ⁺ CD4⁺ T cells in spleen (CD45.2⁺) following co-culture with HI-VSV loaded splenocytes (CD45.1⁺) or HI HSV-2 loaded splenocytes were analysed by flow cytometry. Data are mean \pm s.e.m. * $P < 0.05$;

** $P < 0.001$ (two-tailed unpaired Student's t -test). **b**, **c**, Five weeks after VSV immunization, these mice were challenged intranasally with WT VSV (1×10^7 p.f.u. per mouse). Six days after challenge, VSV-specific antibodies in lysate of brain (**b**) and serum (**c**) were measured by ELISA. Depletion of CD4 T cells was performed on -4, -1, 2 and 4 days after challenge by intravenous injection of anti-CD4 (GK1.5). **d**, Albumin levels in tissue homogenates were analysed by ELISA. Data are mean \pm s.e.m. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$ (Mann-Whitney U -test).

Structural insights into inhibition of lipid I production in bacterial cell wall synthesis

Ben C. Chung^{1*}, Ellene H. Mashalidis^{1*}, Tetsuya Tanino², Mijung Kim³, Akira Matsuda², Jiyong Hong³, Satoshi Ichikawa² & Seok-Yong Lee¹

Antibiotic-resistant bacterial infection is a serious threat to public health. Peptidoglycan biosynthesis is a well-established target for antibiotic development. *MraY* (phospho-MurNAc-pentapeptide translocase) catalyses the first and an essential membrane step of peptidoglycan biosynthesis. It is considered a very promising target for the development of new antibiotics, as many naturally occurring nucleoside inhibitors with antibacterial activity target this enzyme^{1–4}. However, antibiotics targeting *MraY* have not been developed for clinical use, mainly owing to a lack of structural insight into inhibition of this enzyme. Here we present the crystal structure of *MraY* from *Aquifex aeolicus* (*MraY*_{AA}) in complex with its naturally occurring inhibitor, muraymycin D2 (MD2). We show that after binding MD2, *MraY*_{AA} undergoes remarkably large conformational rearrangements near the active site, which lead to the formation of a nucleoside-binding pocket and a peptide-binding site. MD2 binds the nucleoside-binding pocket like a two-pronged plug inserting into a socket. Further interactions it makes in the adjacent peptide-binding site anchor MD2 to and enhance its affinity for *MraY*_{AA}. Surprisingly, MD2 does not interact with three acidic residues or the Mg²⁺ cofactor required for catalysis, suggesting that MD2 binds to *MraY*_{AA} in a manner that overlaps with, but is distinct from, its natural substrate, UDP-MurNAc-pentapeptide. We have determined the principles of MD2 binding to *MraY*_{AA}, including how it avoids the need for pyrophosphate and sugar moieties, which are essential features for substrate binding. The conformational plasticity of *MraY* could be the reason that it is the target of many structurally distinct inhibitors. These findings can inform the design of new inhibitors targeting *MraY* as well as its paralogues, *WecA* and *TarO*.

MraY is a member of the polyprenylphosphate *N*-acetyl hexosamine 1-phosphate transferase (PNPT) superfamily. The PNPT superfamily includes bacterial and eukaryotic integral membrane enzyme families such as *MraY*, *WecA*, *TarO*, *WbcO*, *WbpL*, *RgpG* and *GPT*, which are involved in cell envelope polymer synthesis and protein *N*-linked glycosylation¹. *WecA* and *TarO* are also targets for antibiotic development. *MraY* catalyses the transfer of phospho-MurNAc-pentapeptide from UDP-MurNAc-pentapeptide (UM5A) to the lipid carrier undecaprenyl phosphate (C₅₅-P), yielding undecaprenyl-pyrophosphoryl-MurNAc-pentapeptide, known as lipid I (Extended Data Fig. 1a). This step is essential, rate limiting, and Mg²⁺-dependent¹. It is blocked by five classes of natural nucleoside antibiotics (for example, muraymycin and tunicamycin), and bacteriolytic protein E from bacteriophage ΦX174, with various modes of inhibition^{5–8} (Extended Data Fig. 1a). *MraY*-targeted natural products have gained attention because of their *in vivo* efficacy against pathogenic bacteria including *Mycobacterium tuberculosis*, methicillin-resistant *Staphylococcus aureus* (MRSA), and vancomycin-resistant *Enterococcus* (VRE)^{6,9–12}.

Despite their promise, no antibacterial natural products that target *MraY* have been developed for clinical use, in part owing to a lack of structural information on *MraY* catalysis and inhibition. We carried out structural studies of *MraY* in complex with a naturally occurring inhibitor of *MraY*, muraymycin, which shows antibacterial effects against MRSA, VRE and *Pseudomonas aeruginosa*, and *S. aureus*^{11,13–18}. We used MD2 for our structural, enzymatic and biophysical studies¹⁷ (Extended Data Fig. 1b). The muraymycins are known to be competitive inhibitors for the natural substrate UM5A (ref. 14). Unlike UM5A, the muraymycins do not have pyrophosphate and sugar moieties and have a 5-aminoribosyl group (Extended Data Fig. 1b). Using a radiochemical transfer assay¹⁹, we determined the Michaelis constant (*K*_m) for UM5A with purified *MraY*_{AA} to be ~190 μM (Extended Data Fig. 1c), which is within the range of measured cellular UM5A concentrations²⁰. *MraY*_{AA} activity is markedly reduced after addition of 0.3 μM MD2 (Fig. 1a). Using isothermal titration calorimetry (ITC), we measured the dissociation constant (*K*_d) of MD2 for *MraY*_{AA} to be ~20 nM (Fig. 1b). We generated crystals of *MraY*_{AA} in the presence of MD2, which diffracted to 2.95 Å. Phasing was obtained by molecular replacement using the apo*MraY*_{AA} structure (PDB code 4J72) with all

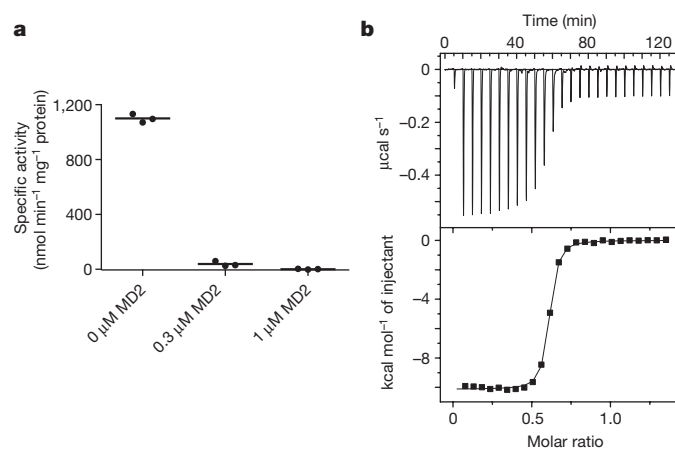


Figure 1 | The natural product MD2 binds to and inhibits *MraY*_{AA}. **a**, Specific activity measurements of wild-type *MraY*_{AA} in the presence and absence of MD2 using radiolabelled substrate, [¹⁴C]UM5A. The radiolabelled product, [¹⁴C]lipid I, was quantified using a liquid scintillation counting method (disintegration per min, d.p.m.). Three technical replicates are shown with the mean value indicated by a line. **b**, Representative ITC raw data (top) and binding isotherm (bottom) for MD2 interacting with wild-type *MraY*_{AA} in the presence of 10 mM MgCl₂; *K*_d = 17.2 nM, Δ*H*^o = −10.1 kcal mol^{−1}. This ITC experiment was performed in triplicate (technical replicates) and mean thermodynamic parameters are shown in Extended Data Table 2.

¹Department of Biochemistry, Duke University Medical Center, 303 Research Drive, Durham, North Carolina 27710, USA. ²Faculty of Pharmaceutical Sciences, Hokkaido University, Kita-12, Nihi-6, Kita-ku, Sapporo 060-0812, Japan. ³Department of Chemistry, Duke University, Durham, North Carolina 27708, USA.

*These authors contributed equally to this work.

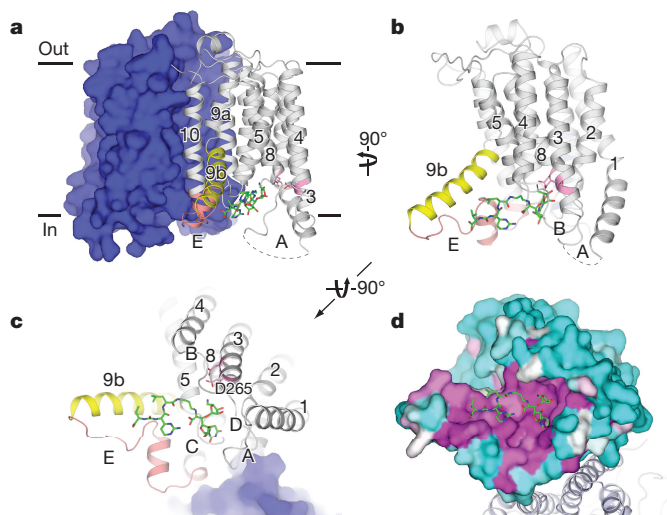


Figure 2 | MD2 binds to a conserved site in MraY_{AA}. **a**, The MD2-bound MraY_{AA} dimer viewed from the membrane. One protomer is shown as surface representation and the other as a cartoon. MD2 (green sticks) resides in the pocket formed by TM3–TM5 and TM8–TM9b and cytoplasmic loops B–E. Conserved catalytic aspartic acid residues at the active site are shown in pink. **b**, View from the membrane rotated 90° about a vertical axis relative to **a**. One protomer is shown for clarity. **c**, Cytosolic view of the MraY_{AA}–MD2 complex, rotated 90° about a horizontal axis relative to **b**. Part of MD2 is near the putative substrate recognition site formed by TM9b (yellow) and loop E (salmon). **d**, Conservation mapping of MraY_{AA} from high (magenta) to low (cyan) sequence identity, based on the alignment of 28 MraY homologues²¹.

the cytoplasmic loops and TM9b removed as a search model. The structure was refined to good statistics ($R/R_{\text{free}} = 0.247/0.261$) (Extended Data Table 1).

MraY_{AA} in complex with MD2 crystallizes as a dimer, as does apoMraY_{AA}²¹. Each protomer contains ten transmembrane helices (TM1–TM10) and five cytoplasmic loops (loops A–E) (Fig. 2a). TM9 breaks into two helical fragments (TM9a and TM9b), and TM9b bends outward towards the membrane (Fig. 2b). We previously outlined the active site as a cleft formed by the inner-leaflet membrane regions of TM3, TM4, TM5, TM8 and TM9b and cytoplasmic loops B, C, D and E. Many absolutely conserved polar/charged amino acid residues are localized in this cleft, including three catalytically critical acidic residues: Asp117, Asp118 and Asp265 (Fig. 2a, c), which are conserved in the PNPT superfamily^{1,19,21–23}. Asp265 interacts with Mg²⁺ in the apoMraY_{AA} structure, and Asp117 has been proposed to bind to the phosphate moiety of C₅₅-P (refs 21, 22). Situated in the active site cleft, the nucleoside portion of MD2 is inserted between loop C and D and the peptide portion interacts with TM9b and loop E (Fig. 2c). High sequence conservation is observed around the MD2-binding region in the active site (Fig. 2d). Notably, MD2 does not interact with any of the three catalytically critical acidic residues (Fig. 2c and Extended Data Fig. 1d).

MraY undergoes remarkable conformational rearrangements near the active site upon binding MD2 (Fig. 3). The amphipathic TM9b rotates away from the active site while loop E rearranges, packs against the hydrophilic part of TM9b (Extended Data Fig. 2 and Supplementary Video 1), and a helical segment of the conserved HHH motif (PXHHHXEXXG) extends^{1,21}. This TM9b–loop E rearrangement widens and reshapes the active site, allowing the peptidic moiety of MD2 to bind to the side of TM9b and loop E (Extended Data Fig. 3). The carboxy-terminal portion of TM5 and loop C unwinds (cyan, Fig. 3a, b) and loop D rearranges (magenta, Fig. 3a, b), creating a pocket where the 5-aminoribosyl moiety and the uracil base of uridine interact (Fig. 3d and Extended Data Fig. 4a–c). The concerted motions of TM5 and loops C and D lead to the rearrangement of loop A and part of TM1

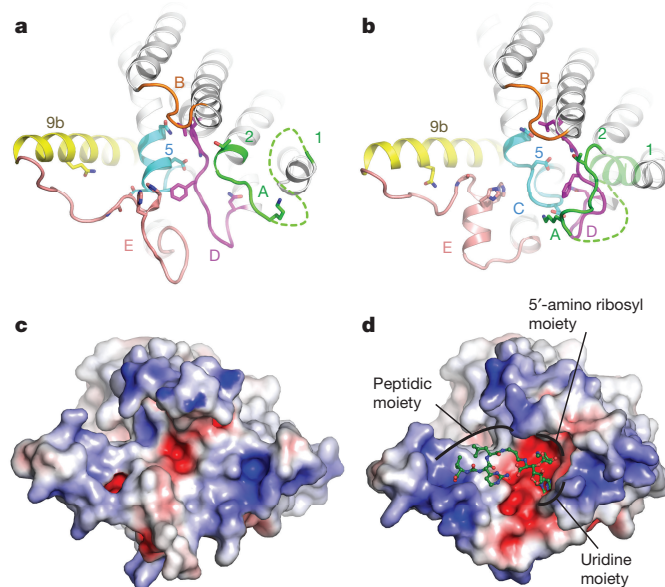


Figure 3 | Conformational rearrangement of MraY_{AA} upon MD2 binding. **a**, apoMraY_{AA} (PDB code 4J72) viewed from the cytoplasm, as in Fig. 2c. Residues involved in interactions with MD2 are shown as sticks. **b**, MD2-bound MraY_{AA} with MD2 omitted. Part of TM1 (light green) is transparent owing to its absence in the apoMraY_{AA} structure. **c**, Electrostatic surface representation of apoMraY_{AA}, viewed from the cytoplasm as in **a** and **b**. **d**, Electrostatic surface representation of MraY_{AA} in complex with MD2. MD2 is green and shown in ball-and-stick representation.

(green, Fig. 3a, b), although these regions do not appear to interact with MD2 directly. It is noteworthy that the amino acid residues interacting with the uracil base move large distances (5–17 Å) while the residues interacting with the 5-aminoribosyl moiety move shorter distances (Extended Data Fig. 4a–c and Supplementary Video 1). The active-site structural rearrangement leads to substantial changes in its electrostatic potential, including enlargement of the acidic milieu around the nucleoside-binding pocket, which may play a role in MD2 binding (Fig. 3c, d).

Developing nucleotide-sugar mimicking inhibitors for glycosyltransferases has been challenging largely owing to the difficulty of developing pyrophosphate mimics capable of cellular entry with high affinity for the target enzyme^{24–27}. MraY is a phosphoglycosyltransferase that shares a common nucleotide sugar substrate with glycosyltransferases. Nucleoside antibiotics targeting MraY have garnered additional interest because they can enter the cell with high affinity for MraY^{24,25,28}. MD2 does not contain a pyrophosphate or sugar moiety and has a 5-aminoribosyl group, which was thought to mimic the pyrophosphate in the natural substrate UM5A (refs 23, 29) (Extended Data Fig. 1b). To understand the principle of MD2 inhibition of MraY_{AA}, we performed mutagenesis and measured the mutational effects on MD2 binding using ITC. Amino acid residues involved in binding MD2 were grouped based on the substructures with which they interact: (1) uracil (Lys70, Gly194, Asp196, Asn255 and Phe262); (2) 5-amino ribose (Thr75, Asn190, Asp193 and Gly264), and (3) peptidic side chain (Gln305, Ala321 and His325). We performed site-directed mutagenesis on the residues that form side-chain interactions with MD2 (Lys70, Thr75, Asn190, Asp193, Asp196, Asn255, Phe262, Gln305 and His325) (Fig. 4). The MraY_{AA} mutants that showed substantial enzymatic activity (Extended Data Fig. 5) were used for ITC experiments with MD2 (Extended Data Fig. 6 and Extended Data Table 2).

The affinity of MD2 for MraY_{AA} was most perturbed with Asp193Asn and Phe262Ala, mutations that disrupt interactions with the 5-aminoribose and uracil moieties of MD2, respectively (Fig. 4, Extended Data Table 2 and Extended Data Fig. 6). Phe262 interacts

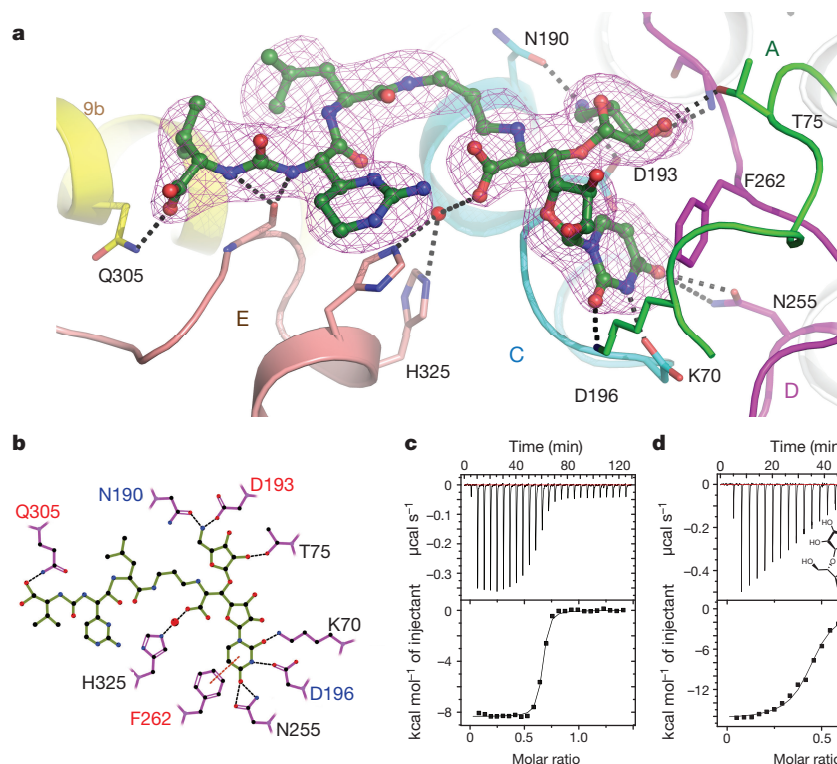


Figure 4 | Dissection of the interactions between MD2 and MrAY_{AA} elucidates the chemical logic of MrAY_{AA} inhibition. **a**, Composite simulated annealing $2F_o - F_c$ omit electron density of MD2 in the crystal structure of MD2-bound MrAY_{AA} at 1.7σ . The transmembrane helices are coloured as in Fig. 3a, b. The residues forming side-chain interactions with MD2 are labelled. **b**, A two-dimensional representation of the interactions between MD2 and MrAY_{AA}. Hydrogen bonds (3.2 \AA cutoff) are indicated with black dashed lines and $\pi-\pi$ contacts are indicated with red dashes. Mutation of residues with red coloured labels resulted in a larger than

fivefold increase in the K_d values of MD2, and those with blue residue labels are nearly inactive. **c**, Representative ITC raw data and binding isotherm for MD2 titrated into MrAY_{AA} in the absence of added Mg^{2+} ; $K_d = 14.8\text{ nM}$, $\Delta H^\circ = -8.3\text{ kcal mol}^{-1}$. A similar K_d value is observed for MD2 titrated into MrAY_{AA} with added Mg^{2+} . **d**, Representative ITC raw data and binding isotherm for 5-aminoribosyl-3-deoxy uridine titrated into wild-type MrAY_{AA}; $K_d = 283\text{ nM}$, $\Delta H^\circ = -16.4\text{ kcal mol}^{-1}$. Each ITC experiment was performed in triplicate (technical replicates) and mean thermodynamic parameters are shown in Extended Data Table 2.

with the uracil base via a $\pi-\pi$ interaction (Fig. 4 and Extended Data Fig. 4d). When Phe262 is mutated to another aromatic amino acid, such as tryptophan, there is a smaller effect on the K_d value relative to the alanine mutation, indicating the importance of this $\pi-\pi$ interaction. Residue Asp193 makes side-chain interactions with the 5-amino ribose moiety of MD2 (Extended Data Fig. 4e). Because the Asp193Ala mutant is nearly inactive (Extended Data Fig. 5b), we used functionally competent Asp193Asn for ITC with MD2 (Extended Data Fig. 5a). However, the heat associated with binding was too low to measure, suggesting that the Asp193Asn mutation greatly reduces the affinity of MD2 for MrAY_{AA} (Extended Data Fig. 6). This observation is consistent with previous studies indicating the antibacterial activity of MrAY inhibitors with a 5-aminoribose is dependent on the amino group of that moiety^{29,30}. The Gln305Ala mutant exhibits a larger than fivefold increase in K_d (Fig. 4 and Extended Data Table 2), indicating that the interactions formed by the peptidic moiety of MD2 contribute to the binding affinity. Asp193, Phe262 and Gln305 are absolutely conserved in MrAY orthologues²¹. The results from the equilibrium binding experiments are consistent with the enzymatic inhibition experiments because the Phe262Ala mutation results in partial inhibition and the Asp193Asn mutant is not inhibited in the presence of $1\text{ }\mu\text{M}$ MD2 (Extended Data Fig. 5a).

We infer that MD2 and the natural substrate, UM5A, use different strategies for binding MrAY. First, the three catalytically critical acidic residues, including the Mg^{2+} -binding Asp265, do not participate in direct interactions with MD2 (Extended Data Fig. 1d). Second, the Asp193Asn mutant remains functionally active, although it disrupts an interaction MrAY makes with the 5-aminoribosyl group and affects the binding affinity of MD2 markedly (Extended Data Table 2 and

Extended Data Fig. 6). This suggests the 5-amino ribosyl group does not function as a pyrophosphate mimic and instead forms interactions that are not present in or important for UM5A binding. If MD2 lacks a pyrophosphate mimic, it is unlikely that Mg^{2+} has an important role in MD2 binding. To test this idea, we performed ITC in the absence of Mg^{2+} and found that MD2 does not require Mg^{2+} for MrAY binding (Fig. 4c). It is possible that the amino group of the 5-aminoribose mimics Mg^{2+} and Asp193 interacts with Mg^{2+} , as previously suggested¹⁴. However, in such a case, we would expect Asp193Asn to be inactive. Furthermore, the amino group of the 5-aminoribose is not near the Asp265, which coordinates Mg^{2+} in the apoMrAY structure²¹.

In summary, the 5-aminoribosyl and uracil moieties of MD2 bind to the nucleoside-binding pocket like a two-pronged electrical plug inserts into a socket, and these interactions are the most critical for binding. The peptidic moiety also contributes to the binding energy by anchoring MD2 to MrAY, probably contributing to its specificity for MrAY. With the 5-aminoribose moiety of MD2 positioned as a second prong alongside uridine in the nucleoside binding site, MD2 binds to MrAY with increased affinity, making the pyrophosphate and sugar moieties unnecessary for binding. To confirm the importance of the 5-aminoribosyl and uracil moieties of MD2, we synthesized 5-aminoribosyl-3-deoxy uridine²⁹ and found it retains substantial binding affinity for MrAY_{AA} (Fig. 4d). Our structural and biochemical studies demonstrate the principles of MD2 inhibition of MrAY and explain why MD2 does not require pyrophosphate and sugar moieties for binding, unlike the natural substrate, UM5A. This illustrates an example of nature circumventing a long-standing problem in chemical biology: developing a nucleotide-sugar-like inhibitor for glycosyltransferases^{24–27}. Finally, the large conformational arrangement observed in MrAY indicates

conformational plasticity, which could be the reason why MraY accommodates so many structurally different nucleoside inhibitors, as well as protein E, with distinct modes of action⁷.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 July 2015; accepted 1 March 2016.

Published online 18 April 2016.

- Bouhss, A., Trunkfield, A. E., Bugg, T. D. & Mengin-Lecreulx, D. The biosynthesis of peptidoglycan lipid-linked intermediates. *FEMS Microbiol. Rev.* **32**, 208–233 (2008).
- Bugg, T. D., Lloyd, A. J. & Roper, D. I. Phospho-MurNAc-pentapeptide translocase (MraY) as a target for antibacterial agents and antibacterial proteins. *Infect. Disord. Drug Targets* **6**, 85–106 (2006).
- Lecerclé, D. *et al.* Bacterial transferase MraY inhibitors: synthesis and biological evaluation. *Bioorg. Med. Chem.* **18**, 4560–4569 (2010).
- Shapiro, A. B., Jahic, H., Gao, N., Hajec, L. & Rivin, O. A high-throughput, homogeneous, fluorescence resonance energy transfer-based assay for phospho-*N*-acetylmuramoyl-pentapeptide translocase (MraY). *J. Biomol. Screen.* **17**, 662–672 (2012).
- Walsh, C. T. & Zhang, W. Chemical logic and enzymatic machinery for biological assembly of peptidyl nucleoside antibiotics. *ACS Chem. Biol.* **6**, 1000–1007 (2011).
- Winn, M., Goss, R. J., Kimura, K. & Bugg, T. D. Antimicrobial nucleoside antibiotics targeting cell wall assembly: recent advances in structure-function studies and nucleoside biosynthesis. *Nat. Prod. Rep.* **27**, 279–304 (2010).
- Brandish, P. E. *et al.* Modes of action of tunicamycin, liposidomycin B, and mureidomycin A: inhibition of phospho-*N*-acetylmuramyl-pentapeptide translocase from *Escherichia coli*. *Antimicrob. Agents Chemother.* **40**, 1640–1644 (1996).
- Bernhardt, T. G., Struck, D. K. & Young, R. The lysis protein E of ϕ X174 is a specific inhibitor of the MraY-catalyzed step in peptidoglycan synthesis. *J. Biol. Chem.* **276**, 6093–6097 (2001).
- Bogatcheva, E. *et al.* Chemical modification of capuramycins to enhance antibacterial activity. *J. Antimicrob. Chemother.* **66**, 578–587 (2011).
- Koga, T. *et al.* Activity of capuramycin analogues against *Mycobacterium tuberculosis*, *Mycobacterium avium* and *Mycobacterium intracellulare* in vitro and in vivo. *J. Antimicrob. Chemother.* **54**, 755–760 (2004).
- McDonald, L. A. *et al.* Structures of the muraymycins, novel peptidoglycan biosynthesis inhibitors. *J. Am. Chem. Soc.* **124**, 10260–10261 (2002).
- Nikonenko, B. V. *et al.* Activity of SQ641, a capuramycin analog, in a murine model of tuberculosis. *Antimicrob. Agents Chemother.* **53**, 3138–3139 (2009).
- Takeoka, Y. *et al.* Expansion of antibacterial spectrum of muraymycins toward *Pseudomonas aeruginosa*. *ACS Med. Chem. Lett.* **5**, 556–560 (2014).
- Tanino, T. *et al.* Mechanistic analysis of muraymycin analogues: a guide to the design of MraY inhibitors. *J. Med. Chem.* **54**, 8421–8439 (2011).
- Tanino, T. *et al.* Synthesis and biological evaluation of muraymycin analogues active against anti-drug-resistant bacteria. *ACS Med. Chem. Lett.* **1**, 258–262 (2010).
- Tanino, T., Ichikawa, S. & Matsuda, A. Synthesis of L-*epi*-capreomycin derivatives via C-H amination. *Org. Lett.* **13**, 4028–4031 (2011).
- Tanino, T., Ichikawa, S., Shiro, M. & Matsuda, A. Total synthesis of (–)-muraymycin D2 and its epimer. *J. Org. Chem.* **75**, 1366–1377 (2010).
- Yamashita, A. *et al.* Muraymycins, novel peptidoglycan biosynthesis inhibitors: synthesis and SAR of their analogues. *Bioorg. Med. Chem. Lett.* **13**, 3345–3350 (2003).
- Lloyd, A. J., Brandish, P. E., Gilbey, A. M. & Bugg, T. D. Phospho-*N*-acetyl-muramyl-pentapeptide translocase from *Escherichia coli*: catalytic role of conserved aspartic acid residues. *J. Bacteriol.* **186**, 1747–1757 (2004).
- Mengin-Lecreulx, D., Flouret, B. & van Heijenoort, J. Cytoplasmic steps of peptidoglycan synthesis in *Escherichia coli*. *J. Bacteriol.* **151**, 1109–1117 (1982).
- Chung, B. C. *et al.* Crystal structure of MraY, an essential membrane enzyme for bacterial cell wall synthesis. *Science* **341**, 1012–1016 (2013).
- Al-Dabbagh, B. *et al.* Active site mapping of MraY, a member of the polyprenyl-phosphate *N*-acetylhexosamine 1-phosphate transferase superfamily, catalyzing the first membrane step of peptidoglycan biosynthesis. *Biochemistry* **47**, 8919–8928 (2008).
- Price, N. P. & Momany, F. A. Modeling bacterial UDP-HexNAc: polyprenol-P HexNAc-1-P transferases. *Glycobiology* **15**, 29R–42R (2005).
- Izumi, M., Yuasa, H. & Hashimoto, H. Bisubstrate analogues as glycosyltransferase inhibitors. *Curr. Top. Med. Chem.* **9**, 87–105 (2009).
- Wang, R. *et al.* A search for pyrophosphate mimics for the development of substrates and inhibitors of glycosyltransferases. *Bioorg. Med. Chem.* **5**, 661–672 (1997).
- Gloster, T. M. & Vocadlo, D. J. Developing inhibitors of glycan processing enzymes as tools for enabling glycobiology. *Nature Chem. Biol.* **8**, 683–694 (2012).
- Rillahan, C. D., Brown, S. J., Register, A. C., Rosen, H. & Paulson, J. C. High-throughput screening for inhibitors of sialyl- and fucosyltransferases. *Angew. Chem. Int. Ed. Engl.* **50**, 12534–12537 (2011).
- Rodolis, M. T. *et al.* Mechanism of action of the uridyl peptide antibiotics: an unexpected link to a protein-protein interaction site in translocase MraY. *Chem. Commun.* **50**, 13023–13025 (2014).
- Dini, C. *et al.* Synthesis of the nucleoside moiety of liposidomycins: elucidation of the pharmacophore of this family of MraY inhibitors. *Bioorg. Med. Chem. Lett.* **10**, 1839–1843 (2000).
- Ii, K., Ichikawa, S., Al-Dabbagh, B., Bouhss, A. & Matsuda, A. Function-oriented synthesis of simplified caprazamycins: discovery of oxazolidine-containing uridine derivatives as antibacterial agents against drug-resistant bacteria. *J. Med. Chem.* **53**, 3793–3813 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements Data for this study were collected at beamlines NE-CAT 24-ID-C and SER-CAT 22-ID-D and at the Advanced Photon Source. We thank K. Yokoyama for advice and guidance throughout the project and Z. Johnson for manuscript reading. Initial X-ray screening of crystals was performed at the Duke macromolecular crystallography facility. This work was supported by NIH R01 GM100984 (S.-Y.L.) and Duke startup funds (S.-Y.L.). This work was also supported by the JSPS Grant-in-Aid for Scientific Research on Innovative Areas ‘Chemical Biology of Natural Products’ (S.I., grant number 24102502) and Scientific Research (B) (S.I., grant number 25293026).

Author Contributions B.C.C. solved the structure and performed some of ITC experiments and E.H.M. carried out the enzymatic assays and performed most of ITC experiments, both under the guidance of S.-Y.L. T.T. carried out chemical synthesis of MD2 under the guidance of S.I. and A.M. M.K. synthesized 5-aminoribosyl-3-deoxyuridine under the guidance of J.H. S.-Y.L., E.H.M. and B.C.C. wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Atomic coordinates and structure factors for the reported crystal structure are deposited in the Protein Data Bank under accession code 5CKR. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.-Y.L. (seok-yong.lee@duke.edu).

METHODS

No statistical methods were used to predetermine sample size.

Crystallization. Wild-type *MraY_{AA}* and mutants were expressed and purified as described²¹. All *MraY_{AA}* bacterial expression plasmids were 10× histidine (His₁₀)-maltose binding protein (MBP) fusion constructs expressed in C41 (DE3) cells. The cells were lysed by microfluidizer and the protein was extracted from the crude lysate using 40 mM dodecyl-maltoside. The lysates were centrifuged to remove the insoluble fraction and the supernatant was applied to a Co²⁺-affinity column for purification. The His₁₀-MBP tag was cleaved overnight by PreScission Protease and *MraY_{AA}* was isolated by gel filtration using a Superdex 200 10/300 GL column in the presence of 5 mM *n*-decyl-β-D-maltopyranoside (DM), 150 mM NaCl, 20 mM Tris-HCl, pH 8.0, and 2 mM dithiothreitol (DTT). All purification steps were performed at 4 °C. After gel filtration, the protein was concentrated to 10 mg ml⁻¹ (~250 μM) and MD2 was added to a final concentration of 400–500 μM before crystallization. Crystals were grown using sitting-drop vapour diffusion in the presence of 50 mM MgCl₂, 40% PEG400, and 100 mM sodium cacodylate, pH 5.6. Crystals were harvested after 10–14 days and flash frozen in liquid nitrogen.

Data collection and structure determination. X-ray data were collected at beamlines 22-ID-D and 24-ID-C at the Advanced Photon Source in Argonne National Laboratory at a wavelength of 1.0 Å and processed using iMosflm. The data were ellipsoidally truncated at 3.0 Å on the *c* axis and anisotropically scaled using the UCLA anisotropy diffraction server (<http://services.mbi.ucla.edu/anisotrope/>). Phases of the MD2 complex structures were solved by molecular replacement using PHASER³¹ with a partial apo*MraY_{AA}* structure (PDB code 4J72, TM9b and cytoplasmic loops removed) as the search model. The omit electron density peaks corresponding to MD2 were prominent from the beginning of the refinement. After the protein model was built, a simulated annealing omit map was generated and MD2 was modelled. PHENIX³² and Coot³³ were used to refine the structure. The final model has good geometry with 98.8%/1.2%/0.0% Ramachandran favoured/allowed/outliers.

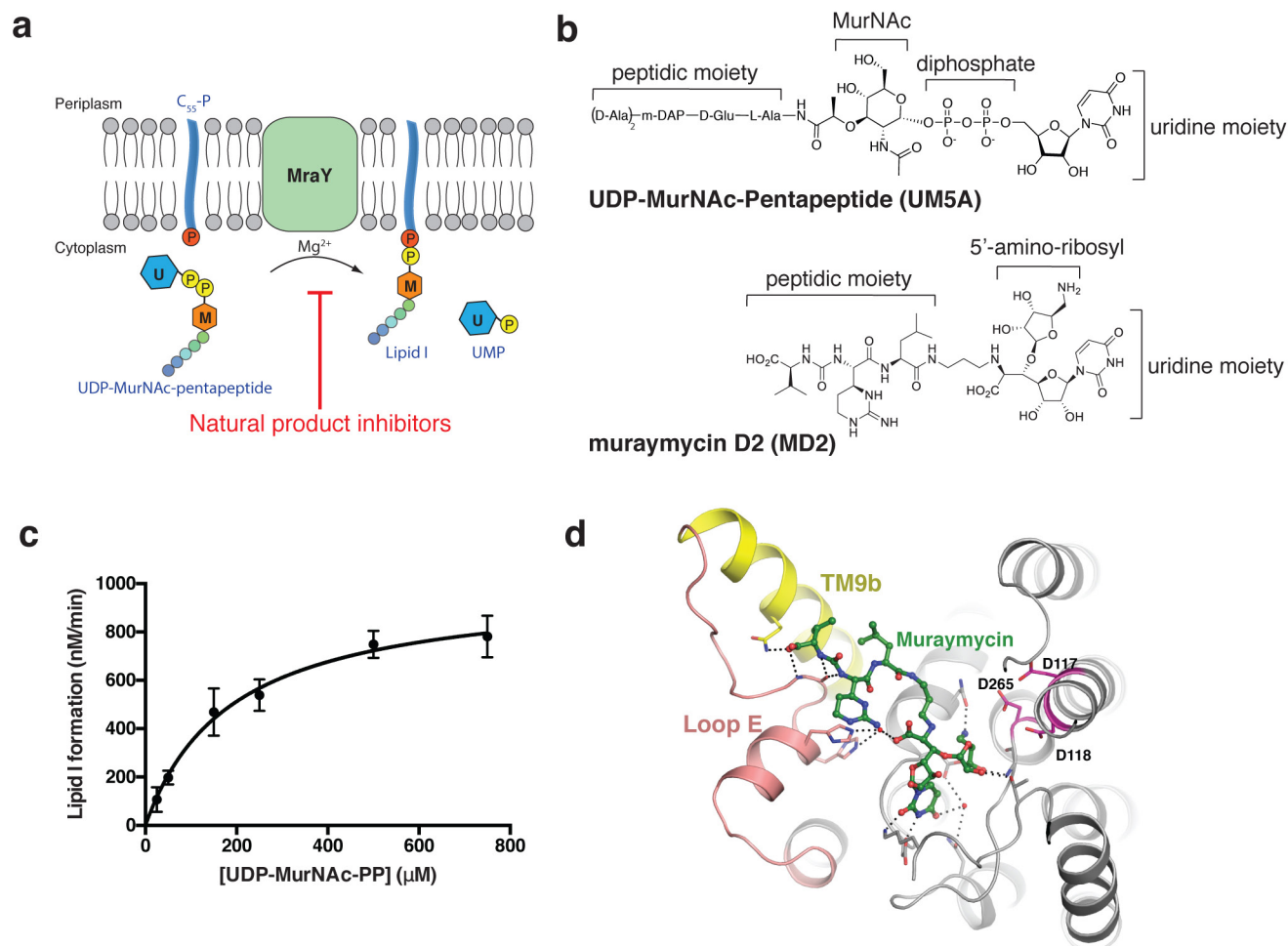
Enzymatic assays. All enzymatic assays monitored the *MraY_{AA}*-mediated transfer of [¹⁴C]phospho-MurNAc-pentapeptide from [¹⁴C]UDP-MurNAc-pentapeptide (DAP) ([¹⁴C]UM5A) to undecaprenyl phosphate (C₅₅-P), forming [¹⁴C]lipid I. A radiochemical transfer assay¹⁹ was used, which was optimized for the *MraY_{AA}*-catalysed reaction as follows. Reaction mixtures (20 μl each) containing 100 mM Tris-HCl, pH 8.0, 500 mM NaCl, 10 mM MgCl₂, 20 mM 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS) and 250 μM C₅₅-P were incubated at 45 °C for 5–6 min in the presence of 0 μM, 0.3 μM or 1.0 μM MD2. For the specific activity assays, 150 μM [¹⁴C]UM5A (3.6 nCi per assay) was used. For wild-type *MraY_{AA}* *K_m* and *k_{cat}* determination, [¹⁴C]UM5A concentration varied from 25 μM to 750 μM (0.6–17.9 nCi per assay). The substrate [¹⁴C]UM5A (specific radioactivity: 1.19 × 10⁻³ Ci mmol⁻¹) was purchased from the BaCWAN facility at the University of Warwick³⁴. Wild-type or mutant *MraY_{AA}* was added to the reaction mixture to a final concentration that enabled product detection within the enzymatic linear range: 50 nM (wild type), 500 nM (Lys70Ala), 400 nM (Thr75Ala), 350 nM (Asp193Asn), 250 nM (Asn255Ala), 200 nM (Phe262Ala), 50 nM (Phe262Trp), 400 nM (Gln305Ala), and 500 nM (His325Ala). The mutant *MraY_{AA}* enzymes Asn190Ala, Asp193Ala, and Asp196Ala were added to the reaction mixture at a final concentration of 500 nM. All wild-type and mutant *MraY_{AA}* enzymes were purified as previously described²¹. Each reaction was initiated with

the addition of enzyme and it was quenched with 20 μl of 6 M pyridinium acetate, pH 3.0. The radiolabelled product, [¹⁴C]lipid I, was isolated from the hydrophilic substrate, [¹⁴C]UM5A, with butanol extraction (200 μl). After vortexing (30 s) and centrifugation at 3,000g (5 min), the upper butanol phase was removed, added to 5 ml scintillation fluid (Fisher Chemical), and analysed using a liquid scintillation counting method (d.p.m.) for ¹⁴C detection (Packard 2500 TR Liquid Scintillation Analyzer). Control reactions lacking enzyme and inhibitor were incubated, extracted and analysed following the same protocol described above and were used for background subtraction. Each reaction rate was calculated by converting the d.p.m. measured (with background subtraction) to moles of [¹⁴C]lipid I formed using the specific radioactivity and dividing by the reaction time. All experiments were performed in triplicate (technical replicates).

Chemical synthesis. MD2 was synthesized as published¹⁷. The MD2 analogue, 5-aminoribosyl 3-deoxyuridine, was synthesized as published³⁵, with a minor modification as follows. In the coupling of the hexose with the ribose, we replaced 5-azido-5-deoxy-D-ribofuranosyl chloride 2,3-diacetate in the original procedure with a synthetically equivalent 5-azido-5-deoxy-2,3-O-(1-methylethylidene)-D-ribofuranosyl fluoride since the ribofuranosyl fluoride was previously prepared in our laboratory³⁶. Except for this minor modification, our synthesis of the MD2 analogue was identical to the original synthesis³⁵.

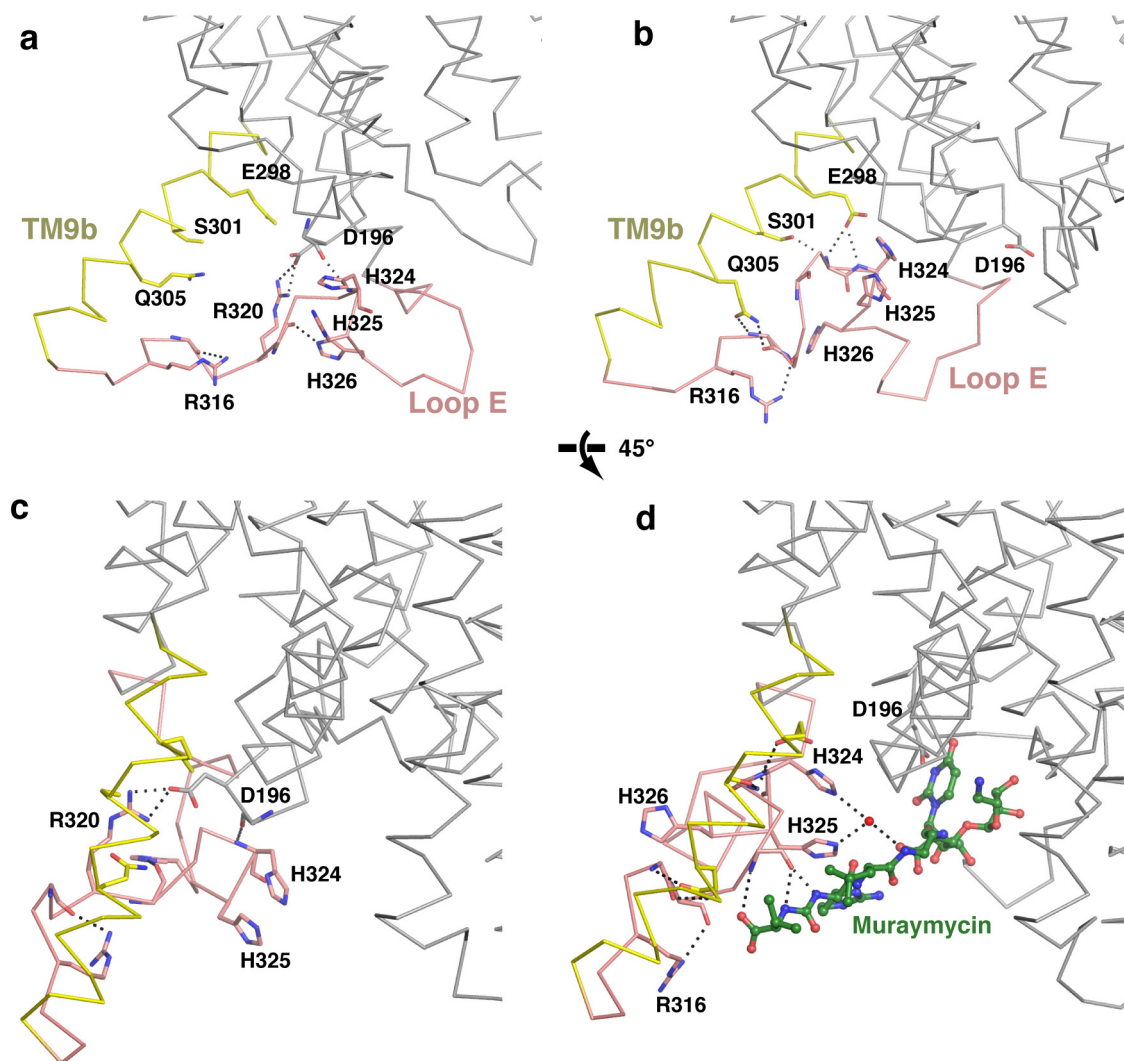
ITC. Wild-type *MraY_{AA}* and mutants (Lys70Ala, Thr75Ala, Asp193Asn, Asn255Ala, Phe262Ala, Phe262Trp, Gln305Ala and His325Ala) were purified as previously described²¹ in a buffer containing 150 mM NaCl, 20 mM Tris-HCl, pH 8.0, 4 mM DM, 2 mM DTT and 10 mM MgCl₂. This same buffer was used to dilute the ligand, MD2. One triplicate set of titrations with wild-type *MraY_{AA}* and MD2 did not include MgCl₂. For wild-type *MraY_{AA}*, 145–240 μM MD2 or 118–130 μM 5-aminoribosyl-3-deoxy uridine was titrated into 6.6–35 μM enzyme. For *MraY_{AA}* mutants Lys70Ala, Thr75Ala, Asp193Asn, Asn255Ala, Phe262Trp and His325Ala, 210 μM MD2 was titrated into 30 μM enzyme. For *MraY_{AA}* Gln305Ala, 315–430 μM MD2 was titrated into 25–27 μM enzyme. For *MraY_{AA}* Phe262Ala, 80–110 μM MD2 was titrated into 7–10.5 μM enzyme. All titrations were performed in triplicate (technical replicates) at 37 °C using either a MicroCal iTC200 or VP-ITC system (GE Healthcare). The total heat exchanged during each injection was fit to a single-site binding isotherm with *K_d* and Δ*H*⁰ as independent parameters. Data were analysed and figures were generated using Origin software (OriginLab Corp).

- McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
- Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
- Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
- Clarke, T. B. *et al.* Mutational analysis of the substrate specificity of *Escherichia coli* penicillin binding protein 4. *Biochemistry* **48**, 2675–2683 (2009).
- Dini, C. *et al.* Synthesis of sub-micromolar inhibitors of *MraY* by exploring the region originally occupied by the diazepanone ring in the liposidomycin structure. *Bioorg. Med. Chem. Lett.* **12**, 1209–1213 (2002).
- Hirano, S., Ichikawa, S. & Matsuda, A. Total synthesis of caprazol, a core structure of the caprazamycin antituberculosis antibiotics. *Angew. Chem. Int. Ed. Engl.* **44**, 1854–1856 (2005).



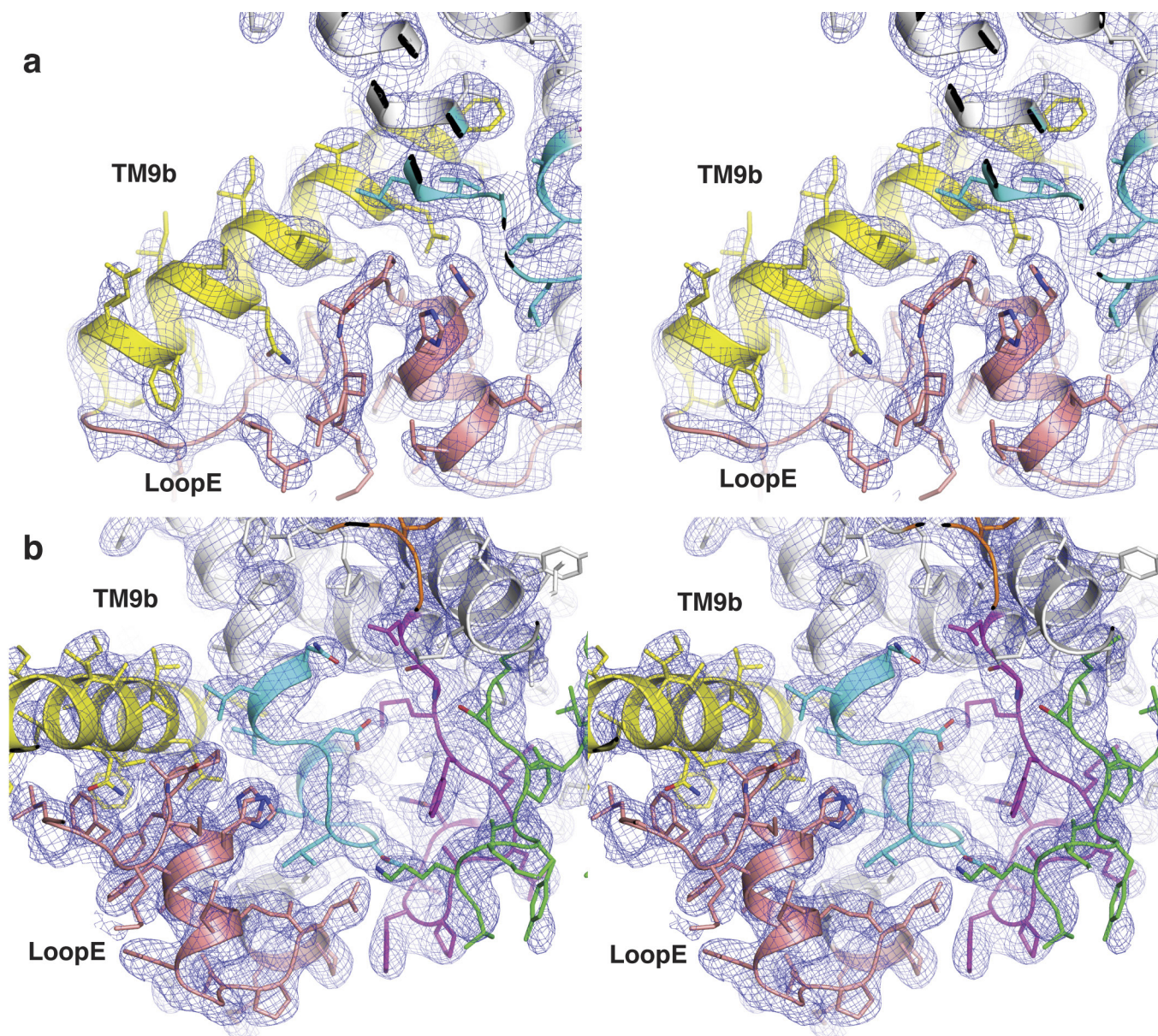
Extended Data Figure 1 | MraY catalyses the formation of lipid I and binds MD2. **a**, Scheme of the reaction catalysed by MraY. The U-labelled blue hexagon represents uridine and the M-labelled orange hexagon represents MurNAc. The phosphates associated with the lipid carrier $C_{55}\text{-P}$ are shown as red circles, and the phosphates from the substrate, UM5A, are shown as yellow circles. **b**, Chemical structures of the substrate, UM5A (top) and the inhibitor MD2 (bottom). **c**, Michaelis-Menten kinetic characterization of MraY_{AA} translocase activity. The reaction monitored

is the MraY_{AA}-catalysed transfer of [^{14}C]phospho-MurNAc-pentapeptide from [^{14}C]UM5A to $C_{55}\text{-P}$, forming [^{14}C]lipid I. The enzymatic parameters measured are as follows: $K_m = 190 \pm 60 \mu\text{M}$, $k_{\text{cat}} = 20 \pm 2 \text{ min}^{-1}$, $k_{\text{cat}}/K_m = 0.11 \pm 0.3 \mu\text{M}^{-1} \text{ min}^{-1}$. Data are mean and s.e.m. of three technical replicates. **d**, MD2 (green) in complex with MraY_{AA}. The distances between MD2 and the three catalytic acidic residues Asp117, Asp118 and Asp265 (magenta) are all greater than 4.5 Å.

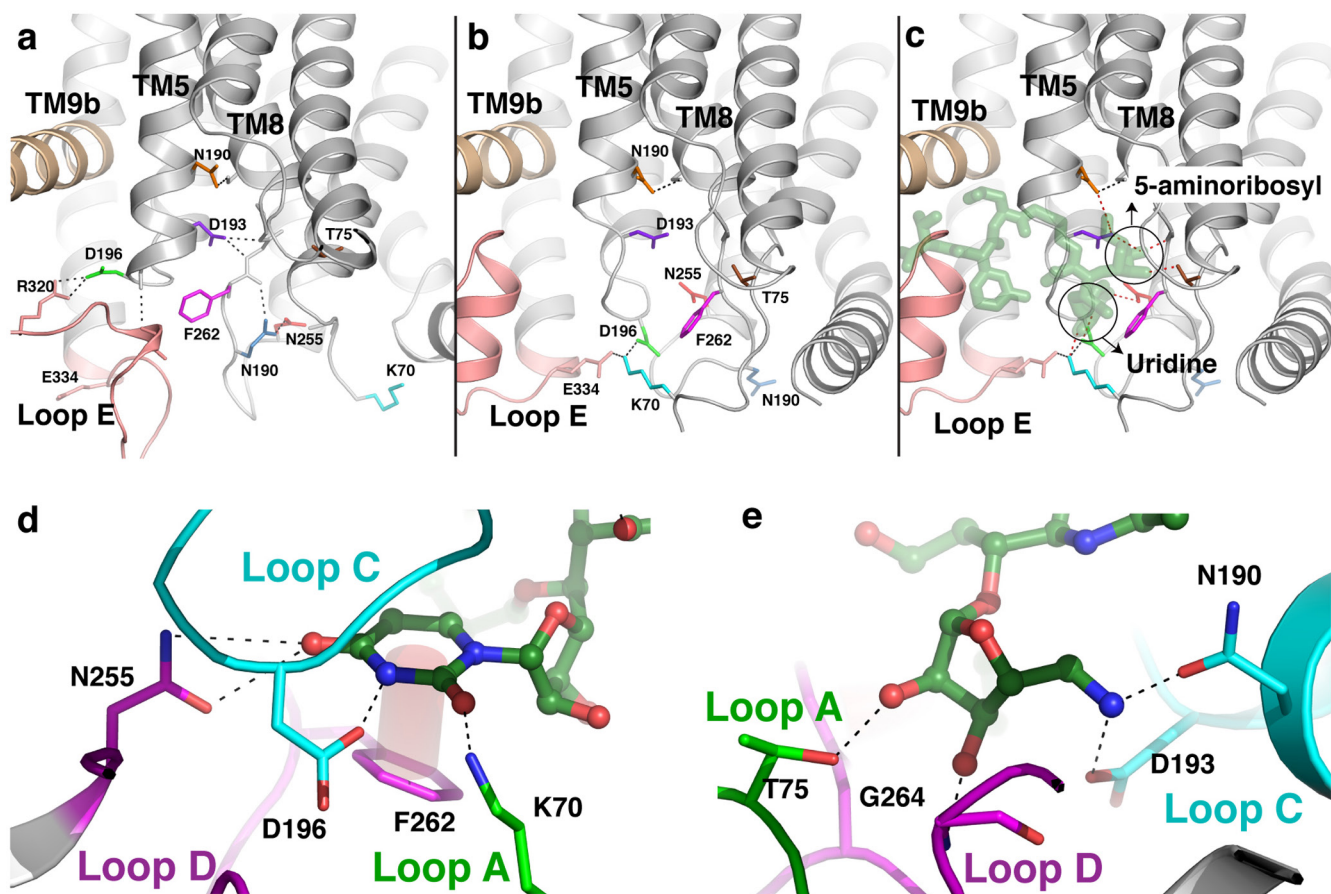


Extended Data Figure 2 | Conformational changes of the TM9b and loop E region of MraY_{AA} upon MD2 binding. **a**, Zoomed-in view of TM9b (yellow) and loop E (salmon) of apoMraY_{AA}, viewed from within the membrane. **b**, TM9b and loop E of MD2-bound MraY_{AA} viewed from within the membrane. MD2 is omitted to illustrate the conformational change of TM9b and loop E associated with MD2 binding.

Conserved amino acid residues and their interactions are shown in stick representation as dotted lines, respectively. **c**, 45° rotated view of **a** about a horizontal axis. **d**, 45° rotated view of **b** about a horizontal axis, including the model of MD2 (green). The rotation of TM9b and rearrangement of loop E, including the HHH motif, allows for MD2 binding, especially its peptidic moiety.

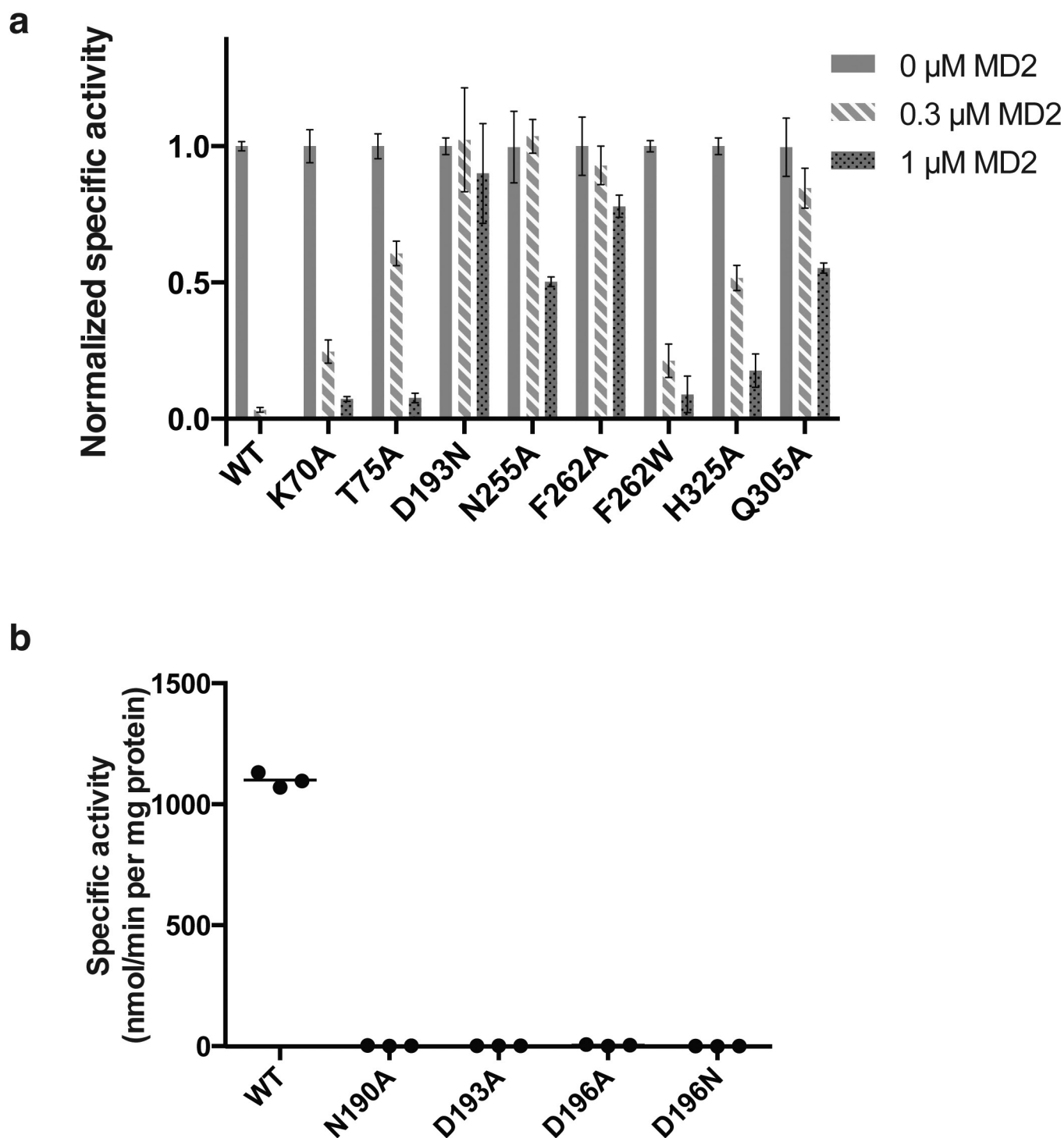


Extended Data Figure 3 | Quality of electron density map surrounding MD2. **a**, Stereo view of $2F_o - F_c$ electron density map at 1σ for TM9b and loop E. **b**, Stereo view of $2F_o - F_c$ electron density map at 1σ for the MD2 binding pocket. The electron density peaks corresponding to MD2 are carved for clarity and all transmembrane helices are coloured as in Fig. 3.



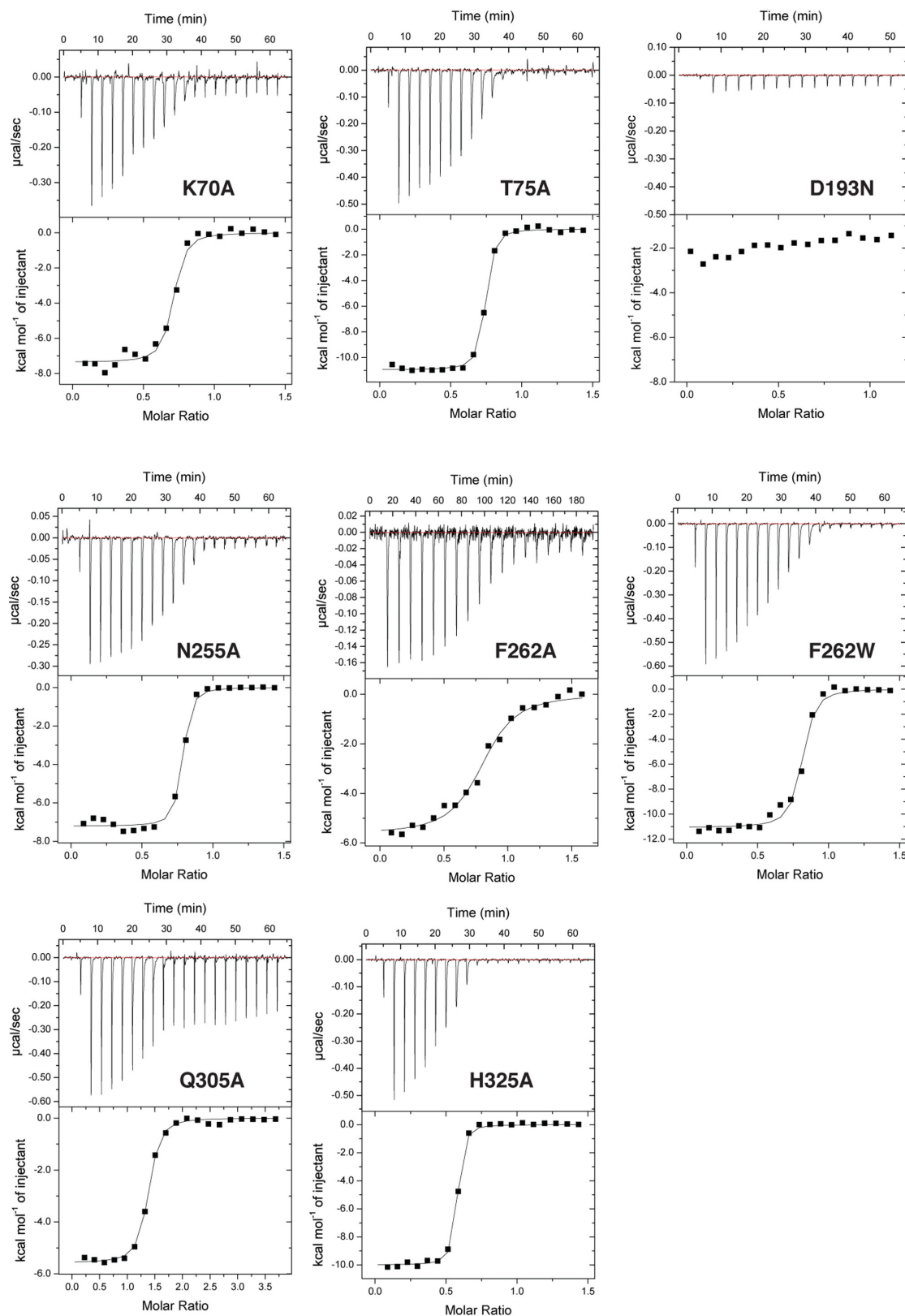
Extended Data Figure 4 | Conformational changes in MraY_{AA} that create binding pockets for the uridine and 5-aminoribosyl groups of MD2. **a**, A close-up view of apoMraY_{AA} with key residues that participate in conformational changes upon MD2 binding shown as sticks in various colours. **b**, A close-up view of the nucleoside-binding pocket in the MraY_{AA}-MD2 complex with MD2 omitted. Key residues are coloured as in **a**. **c**, A close-up view of the interactions MD2 (green) makes with the nucleoside-binding pocket of MraY_{AA}. Interactions between MraY_{AA} and MD2 are shown as dotted lines. It is noteworthy that residues interacting with the uridine moiety of MD2 move large distances (5–17 Å for residues

Lys70, Asp196, Asn255 and Phe262), while the residues binding the 5-aminoribosyl group of MD2 (Thr75, Asn190 and Asp193) do not make large side-chain movements after MD2 binding. The uridine and 5-amino ribosyl groups of MD2 are circled. **d**, Interactions between the uracil base of MD2 (green) and the nucleoside-binding pocket of MraY_{AA}. The uracil base forms H-bonds with side chains of Asn255, Asp196 and Lys70 and forms a π - π interaction with Phe262. **e**, The 5-aminoribosyl group of MD2 forms H-bond interactions with side chains of Thr75, Asn190 and Asp193, and the backbone amide of Gly264.



Extended Data Figure 5 | Specific activity of wild-type and mutant MrayAA in the presence and absence of MD2. **a**, Normalized specific activity of wild-type (WT) MrayAA and enzymatically active mutants with and without MD2 treatment. Wild-type MrayAA or mutant MrayAA was added to the reaction mixture to a final concentration that enabled product detection within the enzymatic linear range: 50 nM (WT), 500 nM (Lys70Ala), 400 nM (Thr75Ala), 350 nM (Asp193Asn), 250 nM (Asn255Ala), 200 nM (Phe262Ala), 50 nM (Phe262Trp), 400 nM (Gln305Ala), and 500 nM (His325Ala). Each reaction was carried out in the presence of either 0 μ M, 0.3 μ M or 1 μ M MD2. Data are shown for three technical replicates \pm s.e.m. Specific activity measurements for each mutant were normalized relative to that without added MD2. **b**, Specific

activity of wild-type MrayAA and enzymatically inactive mutants. MrayAA Asn190Ala, Asp193Ala, Asp196Ala and Asp196Asn were each added to a final concentration of 500 nM, while wild-type MrayAA was present at 50 nM. All enzymatic reactions were conducted with a radiochemical assay monitoring the transfer of [14 C]phospho-MurNAc-pentapeptide from [14 C]UM5A to C₅₅-P, forming [14 C]lipid I. The radiolabelled product, [14 C]lipid I, was quantified using a liquid scintillation counting method (d.p.m.). Specific activity was calculated by determining moles of [14 C] lipid I formed, divided by the reaction time and the quantity of enzyme added. Three technical replicates are shown with the mean value indicated by a line.



Extended Data Figure 6 | Representative ITC raw data and binding isotherms for MD2 interacting with mutant MraY_{AA}. All titrations were performed in triplicate (technical replicates); see source data for all titrations. Representative data are shown. For MraY_{AA} mutants Lys70Ala, Thr75Ala, Asp193Asn, Asn255Ala, Phe262Trp and His325Ala, 210 μ M MD2 was titrated into 30 μ M enzyme. For MraY_{AA} Gln305Ala, 315–430 μ M MD2 was titrated into 25–27 μ M enzyme. For MraY_{AA}

Phe262Ala, 80–110 μ M MD2 was titrated into 7–10.5 μ M enzyme. Mean thermodynamic parameters for triplicate titrations are shown in the Extended Data Table 2. Mean K_d values for each triplicate are as follows: 63.9 ± 4.7 nM for Lys70Ala; 27.4 ± 1.5 nM for Thr75Ala; K_d not determined for Asp193Asn; 29.7 ± 0.8 nM for Asn255Ala; 228 ± 4 nM for Phe262Ala; 68.4 ± 0.9 nM for Phe262Trp; 117 ± 10 nM for Gln305Ala; 24.4 ± 0.4 nM for His325Ala.

Extended Data Table 1 | Data collection and refinement statistics

MraY_{AA}-Muraymycin D2

Data collection	
Space group	C 2 2 2 ₁
Cell dimensions	
<i>a</i> , <i>b</i> , <i>c</i> (Å)	94.48, 102.05, 135.8
α , β , γ (°)	90, 90, 90
Resolution (Å)	2.95 (3.06 – 2.95)*
<i>R</i> _{merge} (%)	27.7 (>100)
<i>R</i> _{pim} (%)	11.6 (64.5)
<i>I</i> / σ <i>I</i>	4.6 (1.2)
CC _{1/2} (%)	87.4 (50.2)
Completeness (%)	100 (100)
Redundancy	7.0 (6.3)
Refinement	
Resolution (Å)	2.95 (3.18 – 2.95) [†]
No. reflections	14053
Completeness (%)	99.40 (93.90)
<i>R</i> _{work} / <i>R</i> _{free} (%)	24.7/26.1
No. atoms	
Protein	2576
Ligand/ion	64
Water	9
<i>B</i> -factors	
Protein	63.50
Ligand/ion	60.50
Water	50.00
R.m.s. deviations	
Bond lengths (Å)	0.008
Bond angles (°)	1.14
Molprobability	
Overall	1.75
Ramachandran (%)	
Favored	99.1
Allow	0.9
Outlier	0

*Values in parentheses are for highest-resolution shell.

[†]Anisotropic truncation at 3.0 Å on *c* axis by the UCLA Diffraction Anisotropy Server (<http://services.mbi.ucla.edu/anisotryscale/>)

Extended Data Table 2 | Equilibrium dissociation constants and binding parameters demonstrating the effect of mutation in MraY_{AA} on MD2 binding

	K_D (nM)	N (sites)	ΔH (kcal/mol)	ΔS (cal/mol/deg)
WT +Mg²⁺	20.4 ± 1.9	0.62 ± 0.02	-15.0 ± 3.5	-11.5 ± 9.8
WT -Mg²⁺	15.1 ± 0.2	0.65 ± 0.05	-9.1 ± 0.6	6.5 ± 2.0
K70A	63.9 ± 4.7	0.72 ± 0.05	-7.2 ± 0.2	9.8 ± 0.7
T75A	27.4 ± 1.5	0.71 ± 0.01	-11.7 ± 0.6	-3.3 ± 1.7
D193N	-	-	-	-
N255A	29.7 ± 0.8	0.68 ± 0.06	-7.6 ± 0.3	10.1 ± 1.1
F262A	228 ± 4	0.77 ± 0.07	-4.9 ± 0.8	14.7 ± 2.5
F262W	68.4 ± 0.9	0.81 ± 0.03	-10.4 ± 0.7	-0.7 ± 2.2
Q305A	117 ± 10	0.73 ± 0.08	-9.8 ± 0.3	0.1 ± 1.0
H325A	24.4 ± 0.4	0.56 ± 0.07	-11.3 ± 0.7	-1.7 ± 2.2
WT + 5-aminoribosyl-3-deoxy uridine	283 ± 3	0.44 ± 0.02	-15.0 ± 1.5	-19.9 ± 4.8

Data are shown as mean and s.e.m. of three technical replicates.

Web
summary

The crystal structure of the MraY enzyme from *Aquifex aeolicus* in complex with the naturally occurring nucleoside inhibitor muraymycin D2 (MD2) reveals that MraY undergoes a large conformational rearrangement near the active site after the binding of MD2, leading to the generation of a nucleoside-binding pocket and a peptide-binding site.

Crystal structure of the human sterol transporter ABCG5/ABCG8

Jyh-Yeuan Lee¹, Lisa N. Kinch^{2,3,4}, Dominika M. Borek^{2,3}, Jin Wang¹, Junmei Wang⁵, Ina L. Urbatsch⁶, Xiao-Song Xie¹, Nikolai V. Grishin^{2,3,4}, Jonathan C. Cohen¹, Zbyszek Otwinowski^{2,3}, Helen H. Hobbs^{1,4*} & Daniel M. Rosenbaum^{2,3*}

ATP binding cassette (ABC) transporters play critical roles in maintaining sterol balance in higher eukaryotes. The ABCG5/ABCG8 heterodimer (G5G8) mediates excretion of neutral sterols in liver and intestines^{1–5}. Mutations disrupting G5G8 cause sitosterolaemia, a disorder characterized by sterol accumulation and premature atherosclerosis. Here we use crystallization in lipid bilayers to determine the X-ray structure of human G5G8 in a nucleotide-free state at 3.9 Å resolution, generating the first atomic model of an ABC sterol transporter. The structure reveals a new transmembrane fold that is present in a large and functionally diverse superfamily of ABC transporters. The transmembrane domains are coupled to the nucleotide-binding sites by networks of interactions that differ between the active and inactive ATPases, reflecting the catalytic asymmetry of the transporter. The G5G8 structure provides a mechanistic framework for understanding sterol transport and the disruptive effects of mutations causing sitosterolaemia.

Cholesterol is an essential component of vertebrate cell membranes. Animals maintain sterol balance by limiting dietary sterol uptake from the gut and promoting sterol secretion from hepatocytes into bile. These physiological processes are mediated by a heterodimeric ABC transporter, consisting of G5 and G8 (refs 1–3) polypeptides, which is embedded in apical membranes of bile ducts and intestinal enterocytes^{4,5}. Mutations in G5 or G8 that block sterol secretion into bile and the gut lumen cause sitosterolaemia, an autosomal recessive disorder in which sterol accumulation leads to premature coronary atherosclerosis.

ABC transporters constitute a ubiquitous protein superfamily that utilizes energy derived from ATP hydrolysis to translocate substrates across membranes⁶. Family members share a common architecture that comprises two transmembrane domains (TMDs) and two nucleotide-binding domains (NBD; specified herein as the contiguous polypeptide domain contributed from each subunit, whereas NBS denotes a composite nucleotide-binding site made up of elements from both subunits). Humans have 48 ABC transporters that are classified into seven subfamilies (A–G)⁷. In the G5 and G8 half-transporters, the NBD is amino (N)-terminal to the TMD, which consists of six transmembrane helices (TMHs) (Fig. 1a). The molecular mechanism by which G5G8 effluxes sterol from plasma membranes remains poorly defined.

Lipid-driven three-dimensional crystallization is a powerful method to determine structures of integral membrane proteins^{8,9}. The only ABC transporters that have been crystallized in lipid bilayers are the bacterial polypeptide processing and secretion transporter (PCAT1, an ABCB homologue)¹⁰ and the maltose transporter–EIIA complex¹¹. No ABCG family member has been structurally characterized. To obtain diffraction-quality crystals, human G5 and G8 were coexpressed in *Pichia pastoris*¹², and tandem affinity chromatography was used to purify stable, monodisperse G5G8 heterodimers that retained ATPase

activity (Extended Data Fig. 1). The protein was reconstituted into di-myristoyl-phosphatidylcholine (DMPC) bicelles¹³, and growth of optimal bicelle crystals required the presence of cholesterol to obtain diffraction higher than 3.9 Å resolution (Extended Data Fig. 2).

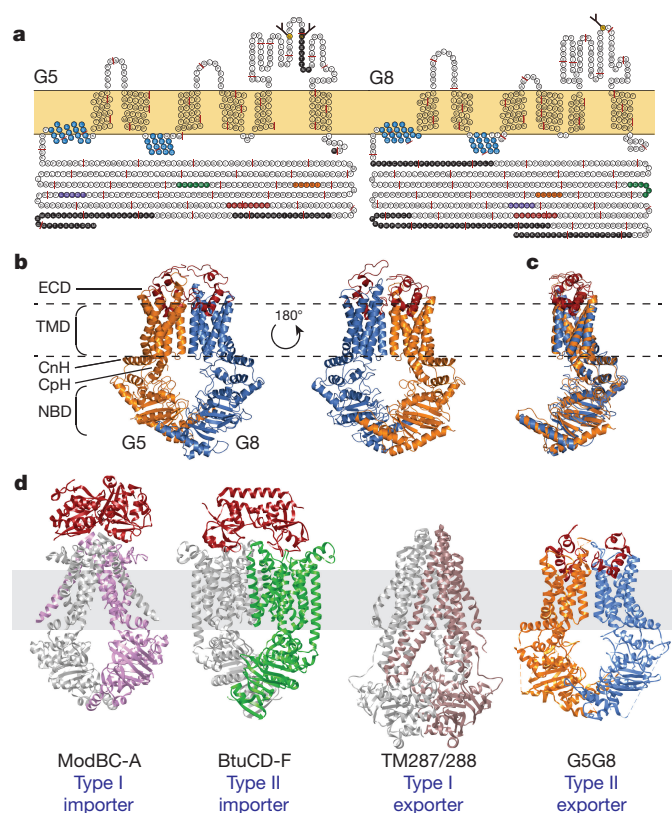


Figure 1 | Structure of the G5G8 heterodimer and comparison with other ABC transporters. **a**, Topological illustrations of G5 and G8: black filled circles for unresolved residues, blue filled circles for the connecting (CnH) and coupling helices (CpH), red filled circles for Walker A motif, green filled circles for Walker B motif, orange filled circles for ABC Signature motif, purple filled circles for the Q-loop. N-linked glycans indicated as branches from asparagine residues. **b**, Cartoon of the G5G8 heterodimer with G5 in orange, G8 in blue, and the ECDs in red. **c**, Superposition of Cα backbones of G5 and G8 (468 out of 571 (G5)/579 (G8) residues). **d**, Nucleotide-free structures of ABC transporter superfamilies: ModBC-A (PDB accession number 2QNK), BtuCD-F (PDB accession number 4Q4H), and G5G8. Substrate binding proteins of BtuF and ModA are coloured red.

¹Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center at Dallas, Dallas, Texas 75390, USA. ²Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. ³Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. ⁴Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. ⁵Cecil & Ida Green Center for Molecular, Computational and Systems Biology, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. ⁶Department of Cell Biology and Biochemistry, Texas Tech University Health Sciences Center, Lubbock, Texas 79430, USA.

*These authors contributed equally to this work.

The G5G8 structure (Fig. 1b) was solved using tungsten-derived single-wavelength anomalous dispersion (Extended Data Table 1 and Extended Data Fig. 3a). The G5G8 crystals comprise two-dimensional layers in which two heterodimers in the asymmetric unit pack in an anti-parallel fashion related by twofold non-crystallographic symmetry (Extended Data Fig. 3b). Native diffraction data were averaged from 19 crystals to 3.9 Å resolution, and the structure was refined to $R/R_{\text{free}} = 0.242/0.328$ (representative electron density maps for selected regions are shown in Extended Data Fig. 3c). G5 and G8 share 28% amino-acid identity, and show a high degree of structural conservation with a root mean squared deviation (r.m.s.d.) of 2.0 Å (Fig. 1c).

Our G5G8 structure adopts an inward-facing conformation, which is analogous to some other nucleotide-free ABC exporters^{10,14,15} and importers^{16,17}. The packing of the TMHs and interfacial contacts of G5 and G8 differ from other ABC transporter structures, including the type I (for example, ModBC-A)¹⁶ and type II (for example, BtuCD-F)¹⁷ importers, and type I exporters (for example, TM287/288)¹⁴ (Fig. 1d). No TMH from either subunit crosses over into the other half-transporter's TMD. In the extracellular domain (ECD), the regions between TMH5 and TMH6 form distinct α -helical structures (Fig. 1b). Three missense mutations causing sitosterolaemia, R419P and R419H in G5 and G574R in G8 (refs 2, 3), are located near the apices of TMH2 and TMH5, respectively (Extended Data Fig. 4a), and both residues are involved in contacts with the ECDs (for example, R419 forms hydrogen bonds with E578 on the G5 ECD). These mutations would be predicted to interfere with the native positions of the ECD helices, suggesting their importance for sterol exit from the TMDs.

The lack of electron density for nucleotide and the spatial separation between opposing Walker A and Signature motifs indicates that the G5G8 structure represents a nucleotide-free state (Extended Data Fig. 4b, c). Nonetheless, the two NBDs contact each other at the extreme cytoplasmic end to form a closed conformation through a pair of NPXDF motifs (G5: NPFDF; G8: NPADF; Extended Data Fig. 4d), which are conserved in the ABCG family and are required for cholesterol efflux by ABCG1 (ref. 18).

How does G5G8 move sterol out of the energetically favourable environment of the plasma membrane? We identified features in our electron density map that may represent cholesterol. These features are located at symmetrical 'vestibules' on opposing faces of the TMD dimer, which open to the bilayer and extend into the centre of the dimer interface (Extended Fig. Data 5a, b). Each vestibule is flanked by TMH1-2 of one TMD and TMH4-6 of the other TMD, with a 'ceiling' formed by an α -helix from the ECD extending into the membrane. Several residues in these vestibules are conserved throughout eukaryotic evolution and may represent binding surfaces or entryways for sterols to access the core of the heterodimer interface. To test this hypothesis, we performed an *in vivo* functional reconstitution assay using adenoviruses to express recombinant G5 and G8 in G5G8 knockout mice^{5,19}. Expression of wild-type (WT) G5 together with WT G8 resulted in a ~30-fold increase in cholesterol transport into bile. In contrast, expression of G8 with G5 containing a substitution (A540F) that occludes the putative cholesterol-binding site failed to restore biliary cholesterol transport despite forming WT levels of the mature G5G8 heterodimer (Extended Data Fig. 5c).

ATP-dependent sterol translocation across membranes requires allosteric communication between catalytic NBSs and substrate-exporting TMDs. In models derived from bacterial transporters, TMD conformational changes result from engagement of the NBDs with 'coupling helices' (CpH) in the intracellular loops of the TMDs²⁰. The TMDs of G5 and G8 each contain a prototypical CpH in the linker between TMH2 and TMH3 that is analogous to the position of coupling helices in other ABC exporters (Fig. 2). Each half-transporter also contains an orthogonal α -helix that we have named the 'connecting helix' (CnH), which is interfacial to the membrane bilayer and connects the NBD to the TMD. The CnH packs against a short cytoplasmic helix containing a conserved glutamate (designated the E-helix). The CpHs

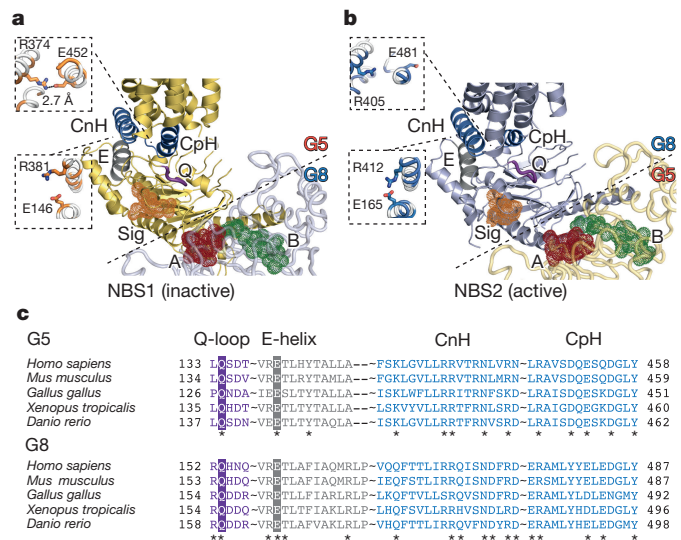


Figure 2 | Interface between NBDs and TMDs of G5 and G8. a, Interface between TMD of G5 and inactive NBS1: connecting (CnH) and coupling (CpH) helices (blue), ABC Signature motifs (orange), Q-loops (purple) and E-helices (grey), Walker A motifs (red) and Walker B motifs (green) of G8. Insets show asymmetric interactions between CnH, CpH, and E-helix. **b**, Interface between TMD of G8 and active NBS2. **c**, Sequence alignments of Q-loop, E-helix, CnH, and CpH: the conserved glutamine of the Q-loop and glutamate in the E-helix are highlighted. Asterisks indicate highly conserved polar residues.

and CnHs are in proximity (~10–15 Å) to the consensus Signature motifs (orange) of the same half-transporters, which form the NBSs with the Walker A (red) and Walker B (green) motifs of the opposing half-transporters (Fig. 2a, b). In G5, R374 on the CnH forms a buried salt bridge with E452 on the CpH, while R381 interacts with the conserved glutamate (E146) that defines the E-helix. A disease-causing missense mutation at this residue (E146Q)³ is predicted to alter crosstalk between the TMD and NBD. The CnH, CpH, and E-helix elements on G5 form a stabilized three-helix bundle. In contrast, the CpH in G8 is rotated such that E481 (corresponding to E452 in G5) points away from R405 (homologous to G5-R374 on the CnH), instead packing against the G8 carboxy (C) terminus that loops back into the heterodimer interface.

The distinct architectures for the two TMD/NBD interfaces reflect the asymmetry in catalytic sites of G5G8. The inactive G5 Signature motif (part of NBS1)²¹, which binds but does not hydrolyse ATP, is adjacent to the CnH/CpH/E-helix bundle of G5, whereas the active G8 Signature motif (part of NBS2) is near the three-helix bundle of G8 (Fig. 2). We hypothesize that the stable three-helix bundle in G5 acts as a rigid body, whereas the three-helix bundle of G8 exhibits greater flexibility and transitions between different conformations. With the proximity of G8-CpH to the catalytically active NBS2, these conformational changes could allosterically link ATP hydrolysis to sterol transport.

What are the likely consequences of ATP binding and hydrolysis on TMD conformation and sterol binding? A network of conserved polar residues in both G5 and G8 forms hydrogen bonds and salt bridges that extend from the CnH and CpH to the proximal part of the TMD interface (Fig. 3). The hydrogen bonds connecting the TMD polar relay may render this network more deformable than a buried hydrophobic core, allowing this region to serve as a flexible hinge for subunit motions with a low energy barrier. Involvement of a TMD polar relay may be a general feature of ABC transporters: the bacterial PCAT1 (ref. 10) exporter and maltose importer²² structures also contain clusters of conserved polar residues in the TMD, which rearrange in the transition to a nucleotide-bound state (Extended Data Fig. 6a, b). We performed a 100-ns molecular dynamics simulation and vibrational mode analysis of our

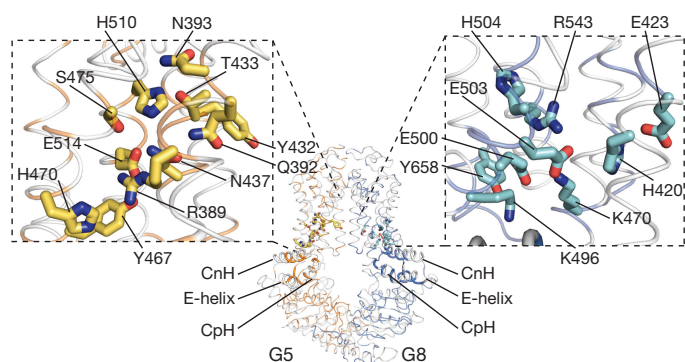


Figure 3 | TMD polar relay of G5 and G8. Stick representation of the conserved polar residues that form the TMD polar relay proximal to CpH and CnH (left: G5; right: G8).

G5G8 structure in an explicit POPC/cholesterol bilayer and water. In these calculations, we found that the CnH and CpH elements of both subunits move inwards, bringing opposing Walker A and Signature motifs into closer contact (Extended Data Fig. 6c). Inward movement of the NBDs was coupled to upward movement of a subset of TMD elements (Extended Data Fig. 6d). R543, which is embedded in the core of the G8 TMD (part of the TMD polar relay), interacts with E503; both residues participate in the upward movement during the molecular dynamics simulation. The sitosterolaemia-causing mutation R543S³ would be predicted to disrupt interaction with E503 and thereby destabilize the TMD polar relay.

We performed coevolution analysis²³ on ABCG TMD sequences to predict positions that could form close contacts at the TMD interface during the transport process (Extended Data Fig. 7a, b). The highest scoring co-evolving positions highlighted three potential interactions between subunits. As these surfaces are distant (>8 Å apart) in the current structure, we infer that they may come into contact at another stage of the transport cycle with inward conformational changes of the TMDs. Y432 on G5 (TMH2) and N568 on G8 (TMH5) are among these pairs, and Y342 is connected to the NBDs through the TMD polar relay (Fig. 3). Accordingly, the mutation G5-Y432A disrupted cholesterol transport in our *in vivo* assay without affecting heterodimer maturation (Extended Data Fig. 7c). Together, these results support a model in which recruitment of ATP stabilizes inward movement of the NBDs and inward/upward movement of the TMDs. We propose that these movements, which would reshape the TMD interface and outer membrane-facing surfaces of the transporter where we observe electron density for cholesterol (Extended Data Fig. 5), contribute to sterol transport across the membrane.

The TMDs of G5 and G8 adopt a tertiary fold that differs from those of known ABC transporter structures²⁰ (Fig. 1d), which have different topologies, longer cytoplasmic extensions of TMH segments and differing placement of coupling helices. Previous classification of ABC exporters suggests that TMDs have arisen at least three independent times, giving rise to the ABC1, ABC2, and ABC3 superfamilies²⁴. G5G8 is a member of the ABC2 exporter superfamily²⁵, which includes the ABCA and ABCG eukaryotic subfamilies and a diverse group of prokaryotic transporters for substrates ranging from polysaccharide-containing teichoic acids and lipo-oligosaccharides (components of bacterial cell walls and outer membranes) to the antibiotic peptides mutacin and bacitracin (Fig. 4a). The ABCG subfamily includes the largest family of ABC transporters in the plant kingdom (Fig. 4b)²⁶ and the historically important pigment transporters that determine eye colour in *Drosophila* (white, brown, and scarlet)²⁷. Using our G5G8 structure as a template, we used the program Modeller²⁸ to create a homology model of the white/brown heterodimer that confers brown eye colour to *Drosophila* (Extended Data Fig. 8). Residues at the end of TMH5 of white (G588, G589, and F590) have been identified as potentially interacting with the dye substrate^{29,30}. These amino acids

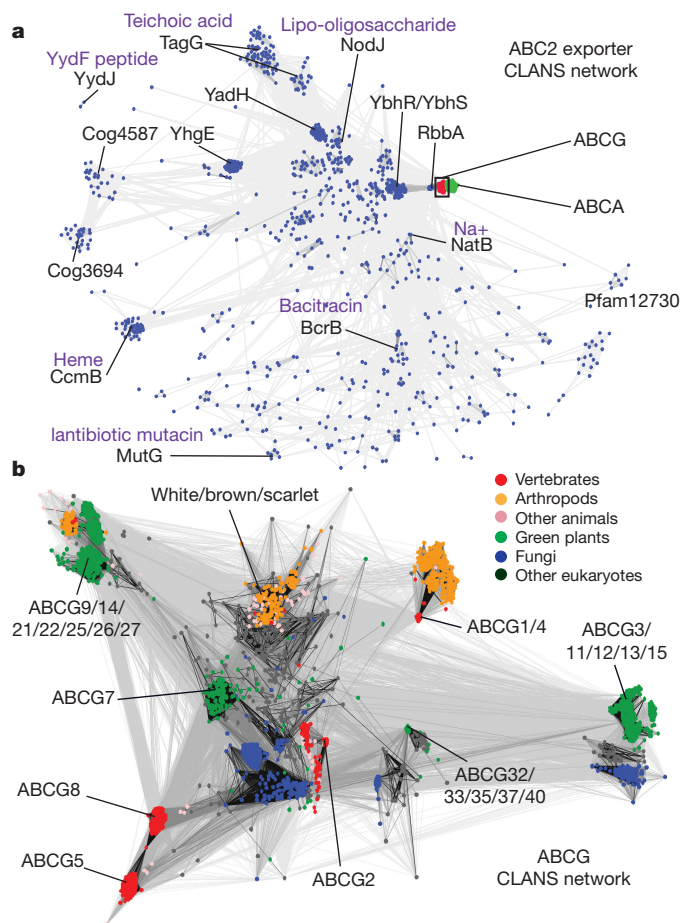


Figure 4 | CLANS network of the ABC2 exporter superfamily and ABCG transporters. **a**, Sequence analysis of the TMD of ABC2 exporter superfamily transporters from both eukaryotes and prokaryotes is shown in a two-dimensional graph output. Lines indicate the similarities between protein sequences with an *E* value < 10^{−10}. Export substrates by prokaryotic transporters in the ABC2 exporter superfamily include polysaccharide teichoic acids, which are found within cell walls of Gram-positive bacteria (TagG); lipo-oligosaccharides, which are located in outer membranes of Gram-negative bacteria (NodJ); haem, which is required for cytochrome C biogenesis (CcmB); and bacterial-killing lantibiotic peptides such as mutacin (MutG) and bacitracin (BcrB). See Methods for more information. **b**, CLANS²⁶ analysis of the TMDs from 4,400 ABCG subfamily members (filled circles) shows functionally diverse clusters.

are conserved in G5G8 at the interface between the TMDs, at a site homologous to the G574R sitosterolaemia mutation in G8. This model indicates a possible conserved site of substrate exit for ABCG transporters and shows that the structure of G5G8 can serve as a powerful platform for structure–function studies of this large and functionally diverse set of membrane proteins.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 14 September 2015; accepted 15 March 2016.

Published online 4 May 2016.

1. Berge, K. E. *et al.* Accumulation of dietary cholesterol in sitosterolemia caused by mutations in adjacent ABC transporters. *Science* **290**, 1771–1775 (2000).
2. Lee, M. H. *et al.* Identification of a gene, ABCG5, important in the regulation of dietary cholesterol absorption. *Nature Genet.* **27**, 79–83 (2001).
3. Lu, K. *et al.* Two genes that map to the STSL locus cause sitosterolemia: genomic structure and spectrum of mutations involving sterolin-1 and sterolin-2, encoded by ABCG5 and ABCG8, respectively. *Am. J. Hum. Genet.* **69**, 278–290 (2001).

4. Graf, G. A. *et al.* Coexpression of ATP-binding cassette proteins ABCG5 and ABCG8 permits their transport to the apical surface. *J. Clin. Invest.* **110**, 659–669 (2002).
5. Graf, G. A. *et al.* ABCG5 and ABCG8 are obligate heterodimers for protein trafficking and biliary cholesterol excretion. *J. Biol. Chem.* **278**, 48275–48282 (2003).
6. Theodoulou, F. L. & Kerr, I. D. ABC transporter research: going strong 40 years on. *Biochem. Soc. Trans.* **43**, 1033–1040 (2015).
7. Dean, M., Rzhetsky, A. & Allikmets, R. The human ATP-binding cassette (ABC) transporter superfamily. *Genome Res.* **11**, 1156–1166 (2001).
8. Caffrey, M. Crystallizing membrane proteins for structure determination: use of lipidic mesophases. *Annu. Rev. Biophys.* **38**, 29–51 (2009).
9. Rosenbaum, D. M. *et al.* GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function. *Science* **318**, 1266–1273 (2007).
10. Lin, D. Y.-W., Huang, S. & Chen, J. Crystal structures of a polypeptide processing and secretion transporter. *Nature* **523**, 425–430 (2015).
11. Chen, S., Oldham, M. L., Davidson, A. L. & Chen, J. Carbon catabolite repression of the maltose transporter revealed by X-ray crystallography. *Nature* **499**, 364–368 (2013).
12. Johnson, B. J. H., Lee, J.-Y., Pickert, A. & Urbatsch, I. L. Bile acids stimulate ATP hydrolysis in the purified cholesterol transporter ABCG5/G8. *Biochemistry* **49**, 3403–3411 (2010).
13. Faham, S. *et al.* Crystallization of bacteriorhodopsin from bicelle formulations at room temperature. *Protein Sci.* **14**, 836–840 (2005).
14. Hohl, M., Briand, C., Grütter, M. G. & Seeger, M. A. Crystal structure of a heterodimeric ABC transporter in its inward-facing conformation. *Nature Struct. Mol. Biol.* **19**, 395–402 (2012).
15. Ward, A. B. *et al.* Structures of P-glycoprotein reveal its conformational flexibility and an epitope on the nucleotide-binding domain. *Proc. Natl Acad. Sci. USA* **110**, 13386–13391 (2013).
16. Hollenstein, K., Frei, D. C. & Locher, K. P. Structure of an ABC transporter in complex with its binding protein. *Nature* **446**, 213–216 (2007).
17. Hvorup, R. N. *et al.* Asymmetry in the structure of the ABC transporter-binding protein complex BtuCD-BtuF. *Science* **317**, 1387–1390 (2007).
18. Wang, F., Li, G., Gu, H.-M. & Zhang, D.-W. Characterization of the role of a highly conserved sequence in ATP binding cassette transporter G (ABCG) family in ABCG1 stability, oligomerization, and trafficking. *Biochemistry* **52**, 9497–9509 (2013).
19. Yu, L. *et al.* Disruption of Abcg5 and Abcg8 in mice reveals their crucial role in biliary cholesterol secretion. *Proc. Natl Acad. Sci. USA* **99**, 16237–16242 (2002).
20. Hollenstein, K., Dawson, R. J. P. & Locher, K. P. Structure and mechanism of ABC transporter proteins. *Curr. Opin. Struct. Biol.* **17**, 412–418 (2007).
21. Wang, J. *et al.* Sequences in the nonconsensus nucleotide-binding domain of ABCG5/ABCG8 required for sterol transport. *J. Biol. Chem.* **286**, 7308–7314 (2011).
22. Oldham, M. L. & Chen, J. Crystal structure of the maltose transporter in a pretranslocation intermediate state. *Science* **332**, 1202–1205 (2011).
23. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife* **3**, e02030 (2014).
24. Wang, B., Dukarevich, M., Sun, E. I., Yen, M. R. & Saier, M. H., Jr. Membrane porters of ATP-binding cassette transport systems are polyphyletic. *J. Membr. Biol.* **231**, 1–10 (2009).
25. Ren, Q., Chen, K. & Paulsen, I. T. TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Res.* **35**, D274–D279 (2007).
26. Frickey, T. & Lupas, A. CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* **20**, 3702–3704 (2004).
27. Morgan, T. H. Sex limited inheritance in *Drosophila*. *Science* **32**, 120–122 (1910).
28. Webb, B. & Sali, A. Protein structure modeling with MODELLER. *Methods Mol. Biol.* **1137**, 1–15 (2014).
29. Ewart, G. D., Cannell, D., Cox, G. B. & Howells, A. J. Mutational analysis of the traffic ATPase (ABC) transporters involved in uptake of eye pigment precursors in *Drosophila melanogaster*. Implications for structure-function relationships. *J. Biol. Chem.* **269**, 10370–10377 (1994).
30. Mackenzie, S. M. *et al.* Mutations in the white gene of *Drosophila melanogaster* affecting ABC transporters that determine eye colouration. *Biochim. Biophys. Acta* **1419**, 173–185 (1999).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the University of Texas Southwestern Structural Biology Laboratory and the staff of the Advanced Photon Source (beamlines 19ID and 23ID) for support during data collection. We thank C. Zelasko, L. Donnelly, F. Xu, L. Nie, Z. Wang, Y. Ma, and C. Zhao for technical assistance. We also thank S. Wilkens for comments, and L. Rice, Y. Jiang, and R. Hibbs for providing reagents and equipment. This project was supported by grants from the American Heart Association South Central Affiliate- (0825285F; J.-Y.L.), the American Heart Association Texas Affiliate Beginning Grant-in-Aid (0463130Y; I.L.U.), the Welch Foundation (I-1770; D.M.R.), the Packard Foundation (D.M.R.), the Howard Hughes Medical Institute (H.H.H., N.V.G.), and the National Institutes of Health (HL72304 and P01-HL20948 (H.H.H., J.C.C., X.-S.X.), GM094575 (N.V.G.), GM053163 and GM117080 (Z.O.), and GM113050 (D.M.R.)). The Advanced Photon Source is a US Department of Energy Office of Science User Facility operated for the Department of Energy Office of Science by Argonne National Laboratory (DE-AC02-06CH11357).

Author Contributions J.-Y.L. and Ji.W. performed the experiments described in this paper with guidance from H.H.H., D.M.R., J.C.C., I.L.U., and X.-S.X. X-ray data were collected by J.-Y.L., and structure determination/refinement was performed by Z.O. and D.M.B. with help from J.-Y.L. L.N.K. and N.V.G. performed bioinformatics and coevolution analysis. Ju.W. performed molecular dynamics simulations and modelling calculations. H.H.H. and D.M.R. supervised the overall project. All authors contributed to the writing of the paper.

Author Information Atomic coordinates and structural factors for the reported crystal structure have been deposited in the Protein Data Bank under the accession number 5D07. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.H.H. (Helen.Hobbs@UTSouthwestern.edu) or D.M.R. (Dan.Rosenbaum@UTSouthwestern.edu).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Cloning and expression of recombinant human ABCG5 and ABCG8. *P. pastoris* expression vectors (pSGP18 and pLIC) were derived from pPICZB (Invitrogen) as described^{12,31}. The cDNAs for human ABCG5 (NCBI accession number NM_022436) and ABCG8 (NCBI accession number NM_022437) were obtained from the National Institutes of Health collection (M. Dean). A tag encoding a rhinovirus 3C protease site followed by a calmodulin binding peptide (CBP) was added to the C terminus of G8 (pSGP18-G8-3C-CBP). A tandem array of six histidines separated by glycine (His₆GlyHis₆) was added to the C terminus of G5 (pLIC-G5-H₁₂). The plasmids were linearized using PmeI and co-transformed into *Pichia* strain KM71H by electroporation. To select for plasmid integration, cells were plated on YPD plates containing sorbitol (1.0 M) and Zeocin (0.5 or 1 mg ml⁻¹), and incubated at 30 °C for 2–5 days. A total of 10–20 yeast colonies from each plate were selected and grown in minimal glycerol yeast nitrogen base (MGY) medium (10 ml), and then induced by adding methanol (0.5%) in minimal methanol (MM) medium. Crude microsomal membranes were prepared, and 30 µg of protein was resolved by SDS–PAGE. Protein expression was analysed by immunoblotting using monoclonal anti-RGSH₄ antibodies (Qiagen) to detect G5 and polyclonal anti-hABCG8 antibodies (see below) to detect G8. The clones expressing the highest level for both G5 and G8 were selected and stored in 15–20% glycerol at –80 °C.

Cell culture and microsomal membrane preparation. A starter yeast culture was prepared by growing transformed yeast in MGY medium (10 ml) to an absorbance $A_{600\text{ nm}}$ of 10. The culture was used to inoculate a litre of MGY medium in a 2.8-l Fernbach flask and grown in a refrigerated Innova shaker (New Brunswick) at 250 rpm for 24 h (28–30 °C). To maximize yield, the acidity of the culture was monitored with the pH maintained at 5–6 using 10% (w/w) ammonium hydroxide (NH₄OH). To induce protein expression, cells were incubated with 0.1% (v/v) methanol for 6–12 h. The methanol concentration was increased to 0.5% (v/v) by adding methanol every 12 h for 36–48 h. Cell pellets were collected and re-suspended in lysis buffer (0.33 M sucrose, 0.3 M TrisCl, pH 7.5, 0.1 M ϵ -aminocaproic acid, 1 mM EDTA, and 1 mM EGTA) to a concentration of 0.5 g ml⁻¹, and stored at –80 °C. Approximately 30 ± 5 g of cell mass was typically obtained from 1 l of cultured cells.

To prepare microsomal membranes, frozen cells were thawed and reducing agent and protease inhibitors were added (final concentration: DTT (10 mM), leupeptin (2 µg ml⁻¹), pepstatin A (2 µg ml⁻¹), and PMSF (2 mM)). Cells were passed through an ice-chilled microfluidizer (Microfluidics) three to five times at 25,000–30,000 psi. The unbroken cells, nuclei and organelles were spun down at 3,500–4,000g for 15 min followed by 15,000g for 30 min, all at 4 °C. Microsomal membrane vesicles were pelleted by ultracentrifugation using a Beckman 45Ti rotor at 40,000–45,000 rpm (maximum 200,000g) at 4 °C for 2 h and re-suspended in buffer A (50 mM Tris pH 8.0, 100 mM NaCl, and 10% glycerol) using a dounce homogenizer, and stored at –80 °C.

Protein purification and pre-crystallization treatment. Frozen microsomal membranes were thawed and the protein concentration was adjusted to 4–6 mg ml⁻¹ using solubilization solution (50 mM Tris-HCl, pH 8.0, 100 mM NaCl, and 10% glycerol, 1% (w/v) β -dodecyl maltoside (β -DDM, Inalco Pharmaceuticals), 0.5% (w/v) sodium cholate (Sigma-Aldrich), 0.25% (w/v) cholesteryl hemisuccinate Tris (CHS-Tris, Anatrace), 5 mM imidazole, 5 mM β -mercaptoethanol (β -ME), 2 µg ml⁻¹ leupeptin, 2 µg ml⁻¹ pepstatin A, 2 mM PMSF). Insoluble membranes were removed by centrifugation using a Beckman 45Ti rotor at 30,000 rpm (100,000g) for 30 min at 4 °C, and a final concentration of 20 mM imidazole and 0.1 mM Tris (2-carboxylethyl) phosphine (TCEP) was added to the solubilized supernatant.

Tandem affinity chromatography was performed. First, the soluble membrane proteins were bound to a nickel-nitrilotriacetic acid (Ni-NTA) column (Qiagen; 1°Ni-NTA) that was pre-equilibrated with buffer A, and washed with 10 column volumes of buffer B (50 mM HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid), pH 7.5, 100 mM NaCl, 0.1% (w/v) β -DDM, 0.05% (w/v) cholate, 0.01% (w/v) CHS (Steraloids), 0.1 mM TCEP) with 25 mM imidazole. The column was then washed with 10 column volumes of buffer B with 50 mM imidazole, and eluted with buffer C (buffer B with 200 mM imidazole, 1 mM CaCl₂, 1 mM MgCl₂). Peak fractions from the Ni-NTA eluates were mixed with equal volume of buffer D1 (buffer B plus 1 mM CaCl₂, 1 mM MgCl₂), and loaded onto a CBP column (Agilent; 1°CBP) that was pre-equilibrated with buffer D1. To exchange the detergent, the CBP column was washed serially with buffer D1 and buffer D2 (buffer B plus 1 mM CaCl₂, 1 mM MgCl₂, 0.1% (w/v) decyl-maltose neopentyl glycol (DMNG, Anatrace), but no β -DDM) in a step-wise fashion: 3 column volumes of D1, 3 column volumes of D1:D2 (3:1, v/v), 3 column volumes of D1:D2 (1:1, v/v),

3 column volumes of D1:D2 (1:3, v/v), and 6–10 column volumes of D2. Finally, the G5G8 heterodimers were eluted with buffer E (50 mM HEPES, pH 7.5, 300 mM NaCl, 2 mM EGTA, 0.1% (w/v) DMNG, 0.05% (w/v) cholate, 0.01% (w/v) CHS, 1 mM TCEP). Divalent cations were added to the CBP eluates to a final concentration of 10 mM MgCl₂ and 10 mM CaCl₂ to quench residual EGTA. No detergent exchange was performed for proteins used for ATPase assays. The N-linked glycans and the CBP tag were cleaved by endoglycosidase H (Endo H, ~0.2 mg per 10–15 mg purified protein) and HRV-3C protease (~2 mg per 10–15 mg purified proteins) for 6–12 h at 4 °C. The CBP tag-free proteins were collected from the flow-through fraction of a second CBP column (2°CBP). G5G8 was concentrated to a volume of 1–2 ml and separated from aggregates, impurities, and enzymes by gel filtration chromatography using an ÄKTA Purifier and a Superdex 200 30/100 GL column (GE Healthcare Life Sciences) in buffer F (10 mM HEPES, pH 7.5, 100 mM NaCl, 0.1% (w/v) DMNG, 0.05% (w/v) cholate, 0.01% (w/v) CHS). The peak fractions were pooled together, and additional HEPES and TCEP were added to final concentrations of 50 mM and 1 mM, respectively.

Purified G5G8 dimers were treated by reductive methylation. Briefly, proteins were incubated twice with 20 mM dimethylamine borane (DMAB) and 40 mM formaldehyde for 2 h at 4 °C on an oscillatory shaker and then 10 mM DMAB was added. After 12 h, the reaction was stopped by 100 mM TrisCl, pH 7.5. For protein relipidation, the methylated proteins were loaded onto a second Ni-NTA column (2°Ni-NTA) that was pre-equilibrated with 100 mM TrisCl, pH 8.0, and 100 mM NaCl. The column was washed slowly with 10 column volumes of buffer G (10 mM HEPES, pH 7.5, 100 mM NaCl, 0.5 mg ml⁻¹ DOPC:DOPE (3:1, w/w), 0.1% (w/v) DMNG, 0.05% (w/v) cholate, 0.01% (w/v) CHS), and the lipidated proteins were eluted using buffer H (10 mM HEPES, pH 7.5, 100 mM NaCl, 200 mM imidazole, 0.5 mg ml⁻¹ DOPC:DOPE = 1:1 (w/w), 0.1% (w/v) DMNG, 0.05% (w/v) cholate, 0.01% (w/v) CHS). TCEP (1 mM) and MgSO₄ (10 mM) were added to the 2°Ni-NTA eluates, and the protein was treated with 5 mM and 2 mM iodoacetamide for 1 h on ice and then passed through a PD-10 desalting column (GE Healthcare Life Sciences) that was equilibrated with buffer I (10 mM HEPES, pH 7.5, 100 mM NaCl, 200 mM imidazole, 0.1% (w/v) DMNG, 0.05% (w/v) cholate, 0.01% (w/v) CHS). Finally, the precipitants were removed by ultracentrifugation using a Beckman TLA120.2 rotor (150,000g) for 10 min at 4 °C. The supernatants were concentrated to a final protein concentration of 25–50 mg ml⁻¹ using a 100 kDa cutoff Vivaspinn concentrator (Sartorius), and used within 1 week of purification for crystallization.

Protein crystallization and crystal sample preparation. All crystals were obtained by reconstituting G5G8 proteins into DMPC/cholesterol/CHAPSO or DMPC/cholesterol/DHPC (Anatrace) bicelles. First, 10% bicelle stock solution was prepared by mixing lipids and detergents (CHAPSO or DHPC) in a ratio of 3:1 (w/w), where the lipids contained 5 mol % cholesterol (Sigma-Aldrich) and 95 mol % DMPC. Immediately before preparing the protein/bicelle mixture, the concentrated proteins were incubated with 10 mM ATP (sodium salt) for 30 min at 4 °C. The proteins and 10% bicelles were then gently mixed in a ratio of 1:4 (v/v), such that the final protein concentration was 5–10 mg ml⁻¹.

The protein/bicelle mixture was incubated on ice for 30 min. The crystallization was set up in a hanging-drop vapour diffusion format at 20–22 °C by using VDX48 trays and mixing protein/bicelle preparation with equal-volume crystallization reservoir solution containing 1.7–2.0 M ammonium sulfate ((NH₄)₂SO₄), 100 mM MES pH 6.5 (or 100 mM HEPES pH 7.0), 2–5% PEG400 (or 2–5% PEG350 MME), and 1 mM TCEP. Crystals suitable for data collection appeared in 3 days to 2 weeks and reached a maximum size of 75–150 µm × 40–60 µm × 10–20 µm in 1–2 months, and decayed after 3 months of crystallization. Crystals used for structural analysis were harvested within 1–2 months. Crystals of similar morphology were also obtained when nucleotide was omitted or when a non-hydrolysable analogue of ATP (either AMPNP or TNP-ATP) was added. We also tried to grow crystals of a catalytically deficient mutant consisting of WT G5 and G8 mutation (G216D)³², attempting to solve a nucleotide-bound structure. No crystal growth was observed, and further optimization in crystallization will be necessary.

For experimental phasing, crystals were derivatized by adding 1 mM sodium phosphotungstate (PW₁₂O₄₀³⁻, Sigma-Aldrich) to the crystallization drops for 12 h before harvesting. Crystals were cryo-protected in a solution containing 25% glycerol, 2 M (NH₄)₂SO₄, 100 mM MES, pH 6.5, and 2–4% PEG400. The crystals were harvested in cryoloops (Mitegen), flash-frozen and stored in liquid nitrogen.

Data collection, structure determination and refinement, final model validation, and uncertainty. X-ray diffraction data sets were collected at the Advanced Photon Source beamlines 19-ID and 23-ID-D. HKL3000 was used to process both the partial and full diffraction data sets used for the ABCG5/ABCG8 heterodimer structure solution^{33,34}. Computational corrections for absorption in a crystal and for imprecise calculations of the Lorentz factor resulting from a minor misalignment of the goniostat were applied^{35,36}. Anisotropic diffraction was corrected to adjust the error model and to compensate for a radiation-induced increase of

non-isomorphism within the crystal^{37–39}. Typical, well-behaving native crystals diffracted anisotropically to a resolution of ~ 3.9 Å in the x direction, ~ 4.0 Å in the y direction, and ~ 4.5 Å in the z direction, while sodium phosphotungstate ($\text{PW}_{12}\text{O}_{40}\text{Na}_3$, Sigma-Aldrich) derivatized crystals diffracted to a resolution of ~ 5.5 Å in the x and y directions, and ~ 6.5 Å in the z direction. The data processing statistics are presented in Extended Data Table 1.

Initial phases were obtained in a single-wavelength anomalous diffraction experiment with two crystals derivatized with sodium phosphotungstate, with data collected at the L-III edge of tungstate ($\lambda = 1.21$ Å). The estimated level of anomalous signal was $\sim 6.5\%$ of the native intensity. The diffraction data set was processed to a resolution of 5.0 Å, while the search for heavy atom positions was performed to a resolution of 7.0 Å. The eight positions of the tungstate cluster were identified using SHELXC/D⁴⁰, run within HKL3000, with correlation coefficients $\text{CC}_{\text{All}} = 45.88\%$, $\text{CC}_{\text{Weak}} = 22.52\%$, and $\text{PATFOM} = 21.56$. The handedness of the best solution was determined with SHELXE, in which the radius of the sphere of the variance map calculation was redefined from 2.42 to 4.84 Å. The eight tungstate cluster positions were refined anisotropically to 5.2 Å with MLPHARE⁴¹, with the final FOM reaching 0.220 for all observations. Twofold NCS was identified by PROFESS from CCP4 (ref. 42) NCS-averaging and solvent flattening was performed by DM⁴³ and later with PARROT⁴⁴. The procedure produced a clean electron-density map to a resolution of about ~ 6.5 Å which showed alpha helical features, but had insufficient resolution to for model building. Therefore, we attempted to improve the phases by combining the phasing signal of the tungstate derivative with the phasing signals of a lead derivative (trimethyl lead chloride, $\text{Pb}(\text{CH}_3)_3\text{Cl}$, Sigma-Aldrich) and tantalum derivative (hexatantalum tetradeabromide, $\text{Ta}_6\text{Br}_{14}$, Jena Bioscience), for which the estimated levels of anomalous signal were below 1% of the native intensity, but which diffracted to better resolution. Three positions of Pb^{2+} were identified using anomalous difference Fourier maps phased with a solvent-flattened tungstate derivative. They were then introduced to MLPHARE together with the previously identified tungstate clusters positions, and refined together to a resolution of 4.2 Å, while the tantalum derivative served as a native data set.

Although the multiple isomorphous replacement phase combination improved the maps, the quality and resolution were still insufficient to build and refine an atomic model automatically. Therefore, we first located the NBD domains using a homologous model (3D31.pdb) and positioned them manually. Then, we placed the alpha helices using the 'Place Helix Here' option in the 'Other Modeling Tools' in Coot⁴⁵. The topology of the transmembrane domains was identified manually and that defined the directionality of the α -helices, which at this resolution were initially placed in two possible directions by Coot. The resulting assembly was used to redefine the solvent mask to improve NCS averaging and then the model was further rebuilt and corrected by iterative application of BUCCANEER⁴⁶, Coot and REFMAC⁴⁷. To proceed with the model building, we had to improve the resolution of the data, which was done by merging together 19 full and partial native data sets to a final resolution of 3.9 Å. Although the data were essentially complete in terms of Bragg's law, the anisotropy of diffraction resulted in an uneven distribution of information in the reciprocal space. Correcting for anisotropy retained informative observations of intensity in ellipsoidal resolution shells; however, all downstream procedures reported completeness in the spherical shells. Therefore, the completeness in refinement is lower than in scaling. The merged native data set was combined with the PARROT-modified phases obtained during an earlier step and was used in the model building with BUCCANEER and in the refinement with REFMAC. The initial model served as the starting point for BUCCANEER, which rebuilt it to $\sim 65\%$ of completeness and partly assigned it to the sequence, with $\sim 60\%$ of its side chains docked. Intermediate models from different cycles of BUCCANEER were combined into a more complete model and rebuilt manually. The restrained, phase-stabilized refinement was performed with the Hendrickson–Lattman coefficients from PARROT down-weighted by factor of 0.5 and blurred with a B-factor of 200 Å². Additionally, the ProSMART⁴⁸ option 'to generate H-bond restraints (e.g. secondary structure restraints)' was used to stabilize the model geometry during refinement, together with REFMAC's local NCS restraints, jelly body refinement, and B-factor values restrained with a weight of 0.2 and to a range of allowable values 20 – 400 . The resolution cutoff of 3.94 Å in the refinement was selected on the basis of multiple refinements in which the R_{free} values in the last resolution shell were inspected and the features of electron density maps were analysed. All these parameters were chosen to stabilize the refinement and reduce bias. The model quality was validated with Molprobity⁴⁹ and assessed as satisfactory with a Molprobity score of 3.47 , which corresponds to the 71st percentile in comparison with the set of 342 Protein Data Bank deposits solved at resolutions from 3.25 to 4.18 Å.

Although nominal resolution of diffraction data extends to 3.94 Å, high anisotropy results in uneven quality of maps in real space, with some electron density features being very well defined while other features have less definition.

Specifically, the amino-acid register of the NBDs, having predominantly β -sheet structure, is highly uncertain when based on electron density alone. We used Robetta models⁵⁰, sequence homology, and analysis of plausibility of chemical interactions to validate the amino-acid register, but we expect that registry errors are possible in these domains, in particular in regions that do not have reliable alignments to homologous structures. In the case of transmembrane domains, we do not expect major registry problems; however, minor mistracings of loops are possible owing to an uncertainty that is related to the C-caps of helices, which can have multiple alternative conformations that are not discernible at this resolution. Finally, although the fragment that connects the NBD with the transmembrane domains (from residues ~ 320 to ~ 395 both in ABCG5 and in ABCG8) can be traced for ABCG5 (chains A and C), it is less well defined in the ABCG8 subunits (chains B and D). For chain B, a part of this fragment is stabilized in the α -helical conformation by crystal lattice interactions, but the linker sequences connecting this fragment to the NBD on one side and the transmembrane domain on the other are not visible in the electron density, and therefore the register for this fragment is highly uncertain.

ATPase assay. The ATPase activity of purified G5G8 was determined as described^{12,51}. Briefly, 4 – 10 μg proteins (final concentration 27 – 67 $\mu\text{g ml}^{-1}$) were mixed with 100 μg liver polar lipids (Avanti) and 20 mM DTT for 10 min at room temperature (22°C), and then left on ice (used within 2 h). Reactions were performed in a final volume of 150 μl containing 50 mM Tris/MES (pH 7.0), 30 mM KCl, 5 mM MgSO_4 , 2 mM ^{32}P - γ -ATP, 4.5 mM sodium azide, and 1% sodium cholate at 37°C . Released inorganic phosphate was extracted by molybdate and monitored by ^{32}P radioactivity to measure specific activity in three independent experiments. As a negative control, we paired a catalytically deficient G8 mutation (G216D)³² with WT G5 in the assay.

Generation of anti-human G5 and G8 antibodies. A DNA fragment encoding the N-terminal region of human G8 (residues 2 – 400) was PCR-amplified from the G8 cDNA and cloned into the pET30a⁺ vector (Novagen). The peptide was expressed in BL21 (DE3) competent *Escherichia coli* cells (Novagen) and then isolated from inclusion bodies, solubilized in 8 M urea, and purified using a Ni-NTA column in the presence of 8 M urea. Rabbits were injected every 2 weeks with 0.1 mg of the peptide to generate anti-G8 polyclonal antibodies. To generate monoclonal anti-human G5 antibodies, splenic B lymphocytes were isolated from female BALB/c mice ($n = 2$) that had been immunized eight times with 50 μg of purified G5G8 (see above). Cells were incubated with SP2/mIL-6 mouse myeloma cells and the resultant hybridomas were screened using an ELISA assay. Positive clones that recognized denatured G5 by immunoblotting were selected for subcloning. Class I immunoglobulin- γ (IgG1) was purified from the supernatant of cultured hybridoma cells using gravity-flow affinity chromatography with Protein-G Sepharose-4 Fast Flow beads.

In vivo functional reconstitution cholesterol transport assay. Point mutations were introduced into the human G5 cDNAs using QuickChange II site-directed mutagenesis kits (Agilent). The recombinant adenoviruses expressing human WT or mutant G5 and G8 were generated using an AdenoVator adenoviral vector system (QBioGene). Total knockout (KO) ($\text{Abcg5/Abcg8}^{-/-}$)¹⁹ and liver-specific KO ($\text{L-Abcg5/Abcg8}^{-/-}$)⁵² mice were maintained on a regular chow diet. Adenoviral particles (5×10^{12} particles per kilogram), containing no external gene (RR5), WT, or mutant human G5G8, were injected into the tail veins of the mice. After 72 h, the mice were fasted for 4 h, anaesthetized with halothane, and killed by exsanguination. Bile was collected, and neutral sterol levels were measured using gas liquid chromatography and mass spectrometry as described⁴. Liver tissue was snap frozen in liquid nitrogen and stored at -80°C .

G5G8 coevolution analysis. Information about the covariance of residue substitution patterns in multiple sequence alignments (MSAs) or co-evolution can provide a basis for predicting structure contacts^{53,54}, and has been successfully applied to predicting residue–residue interactions across protein interfaces²³. We employed the GREMLIN server⁵³ to predict the top $L/3$ co-evolving residue pairs in the TMD of ABCG5 and ABCG8 (see Supplementary Information). We mapped the co-evolving pairs that were separated by at least four residues in the primary sequence to the corresponding structures and compared them with the ABCG5 and ABCG8 TMD structure residue–residue contact maps using CMview⁵⁵. GREMLIN co-evolving residue pairs from the ABCG5 and ABCG8 TMD that were closer than an all-atom cutoff of 8 Å represent the majority of co-evolving residue pairs and are depicted as lines connecting C α atoms of the residues in contact in the respective TMD structures. The remaining residues that are spatially distant within the G5 or G8 molecules are candidates for forming interfacial contacts between G5 and G8 (ref. 23). For those remaining Gremlin pairs in TMH1, TMH2, and TMH5, we used the Gremlin alignment to map residues from G5 (or G8) to the corresponding residues on the opposite molecule. The highest scoring interfacial Gremlin pairs (three pairs with > 0.95 probability of co-evolving, corresponding to six potential symmetry-related pairs) were between residues in TMH2 (G8 L459,

F461, and Y465 or the corresponding G5 T430, Y432, and L436) and residues in TMH5 (G5 A535, I539, S538 or the corresponding G8 N564, N568, and Y567, respectively) (Extended Data Fig. 7).

Molecular dynamics simulation. CHARMM-GUI (<http://www.charm-gui.org>) was applied to add DMPC lipid bilayer, cholesterol, counter ions, and water molecules. The entire system consisted of one copy of the G5G8 biological unit (chains A and B), 16 cholesterol (CHL), 304 DMPC, 128 Na⁺, 180 Cl⁻, and 47,789 TIP3P water molecules⁵⁶. In total, there were 163,088 atoms in the simulation box for the whole system. For force field parameters, the partial atomic charges of CHL were derived by RESP⁵⁷ to fit the HF/6-31G electrostatic potentials generated using the GAUSSIAN 09 software package (revision D.01). The other force field parameters came from GAFF in AMBER12 (ref. 58). The residue topology of CHL was prepared using the ANTECHAMBER module in AMBER12 (ref. 59). The force fields of AMBER FF12SB⁶⁰ and LIPID14 (ref. 61) were used to model proteins and lipids, respectively.

All molecular dynamics simulations were performed with periodic boundary condition to produce isothermal–isobaric ensembles using the PMEMD.CUDA program in AMBER12. The particle mesh Ewald method⁶² was used to calculate the full electrostatic energy of a unit cell in a macroscopic lattice of repeating images. All bonds were constrained using the SHAKE algorithm⁶³ in molecular dynamics simulations. Temperature was regulated using Langevin dynamics⁶⁴ with a 5 ps⁻¹ collision frequency. Pressure was regulated using the isotropic position scaling algorithm with the pressure relaxation time set to 1.0 ps. The integration of the equations of motion was conducted at a time step of 1 fs for the relaxation and equilibrium phases and 2 fs for the sampling phases. Before molecular dynamics simulations, the systems were relaxed to remove any possible steric clashes by a set of 10 thousand-step minimizations with the main chain atoms restrained. The harmonic restraint force constants decreased from 20 to 10, 5, and 1 kcal mol⁻¹ Å⁻², progressively. At last, the systems were further relaxed by a 10,000-step minimization without any constraint or restraint. There are three phases in a molecular dynamics simulation: the relaxation phase, the equilibrium phase, and the sampling phase. In the relaxation phase, the system was gradually heated up from 50 K to 300 K in steps of 50 K. At each temperature, molecular dynamics simulation was run for 1 ns. In the following equilibrium phase, the system was further equilibrated for 2 ns at 298 K. In the sampling phase, 50,000 snapshots were collected at an interval of 2 ps for post-analysis. All the post-analysis was performed using the CPPTRAJ module of AMBER12.

The stability of the molecular dynamics trajectory was monitored with the backbone r.m.s.d. along the simulation trajectory, showing that the current crystal structure was very stable along the whole simulation course with an overall r.m.s.d. of 3.0 Å, which is reasonable for a structure resolved at 3.9 Å. A subset of 10,000 snapshots was evenly selected for the solvent-accessible surface calculation and cluster analysis. We found no obvious trend for the solvent-accessible surface areas of TMHs. We then performed fixed radius clustering analysis (radius was set to 2.0 Å) for the subset using the clustering toolkit of MMTSB (<http://www.mmtsb.org>). The biggest cluster, 5,793 members and the shortest distance to the cluster centre, was selected to compare with the crystal structure. Only slight narrowing in the TMD could be observed, but the NBD had larger deviations, probably owing to the missing residues not resolved in this study.

The quasi-harmonic analysis was performed for the main chain atoms using all 50,000 snapshots collected every 2 ps. First, an average molecular dynamics structure was calculated by aligning the molecular dynamics snapshots to the crystal structure using the main chain atoms; each snapshot was then realigned to the average structure to generate the mass-weighted covariance matrix. Vibrational frequencies and modes, which are the eigenvalues and eigenvectors of the covariance matrix, were then obtained. For each vibrational mode, the contribution of each residue was calculated by adding up the vectors of the main chain atoms. The low-frequency modes usually correspond to the global movements of a protein related to its biological function⁶⁵. Among the lowest 20 vibrational modes, modes 9 and 10 were selected for further examination (Extended Data Fig. 6). Under current quasi-harmonic analysis, we were not able to observe a mode describing opening or closing at the TMD interface. Experimental structures of different catalytic or substrate-bound states will be necessary.

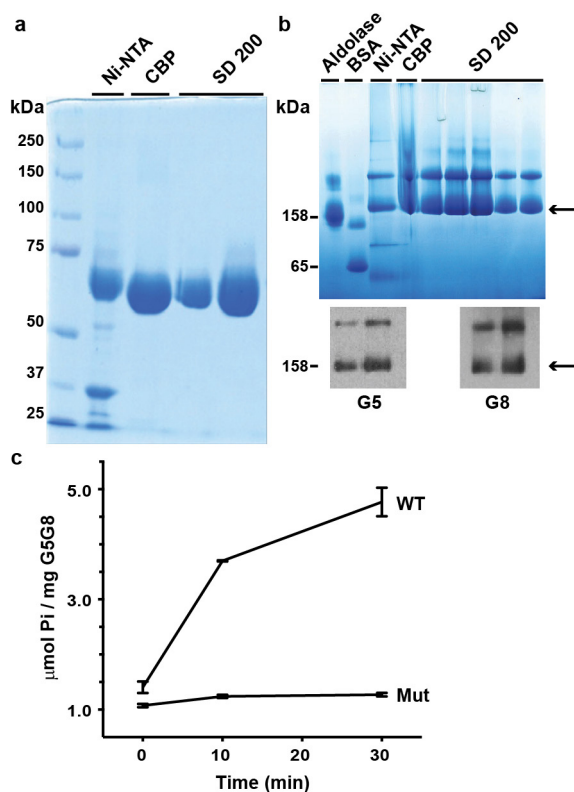
Homology modelling. The homology model of the white/brown heterodimer that confers brown eye colour to *Drosophila* was generated using Modeller²⁸ with the G5G8 crystal structure as the template. The sequence alignment for homology modelling was generated by PROMALS3D⁶⁶. One hundred homology models were generated and the one with the best DOPE score was selected as the final model. The final model was further evaluated using the PROTABLE module of SYBYL software (<https://www.certara.com>) and no severe violation was identified in the Ramachandran plot.

G5G8 sequence analysis. To collect ABCG family sequences, we ran PSI-BLAST (three iterations, *E* value cutoff 0.001) against the NCBI NR database using the G5 TMD as a query sequence (gi|11692800, residue range 368–651). Collected sequences were clustered using CLANS²⁶. Sequences that did not cluster with the eukaryotic ABCG family were excluded to make a subset of ABCG sequences. To establish residue conservations for the G5 and G8 subfamily subsets, sequences from our initial PSI-BLAST against the NR database that clustered together with the human G5 and G8 sequences were kept, and an MSA was generated for each subfamily using MAFFT⁶⁷. MSA results of G5 and G8 are shown in the Supplementary Information. Redundant and incomplete sequences were manually removed from the MSA. The MSA was used to generate positional conservations for each subfamily and they were mapped to the last line of the MSA using a scale ranging from variable to conserved (scale 0–9), to the structure B-factors in G5 (scale –1.82 to 1.52), and in G8 (scale –2.04 to 1.41) using the program AL2CO⁶⁸. Residues with conservation values of 0.6 or greater in the B-factors (MSA conservation value ≥ 7) were considered highly conserved.

To gain a broader view of all sequences related to the helical domain of TMD, we collected all related eukaryotic sequences by searching the SWISSPROT sequence database with the same G5 query. We also collected related prokaryotic sequences by initiating PSI-BLAST with representative queries from all COGS that belong to the ABC2 exporter superfamily clan (COG1277, COG1682, COG2386, COG559, COG3694, COG4200, and COG4587). To cut down on redundancy of the NR database, we searched a subset of 122 high-quality curated bacterial genomes designated by NCBI as reference genomes (<http://www.ncbi.nlm.nih.gov/genome/browse/reference/>). The sequences, identified in the SWISSPROT database, were combined with those identified in the prokaryotic reference genomes and were clustered in two dimensions using CLANS²⁶ to visualize the ABC2 exporter superfamily sequence relationships. Export substrates by prokaryotic transporters in the ABC2 exporter superfamily include polysaccharide teichoic acids, which are found within cell walls of Gram-positive bacteria (TagG)⁶⁹; lipo-oligosaccharides, which are located in outer membranes of Gram-negative bacteria (NodJ)⁷⁰; haem, which is required for cytochrome C biogenesis (CcmB)⁷¹; and bacterial-killing lantibiotic peptides such as mutacin (MutG)⁷² and bacitracin (BcrB)⁷³.

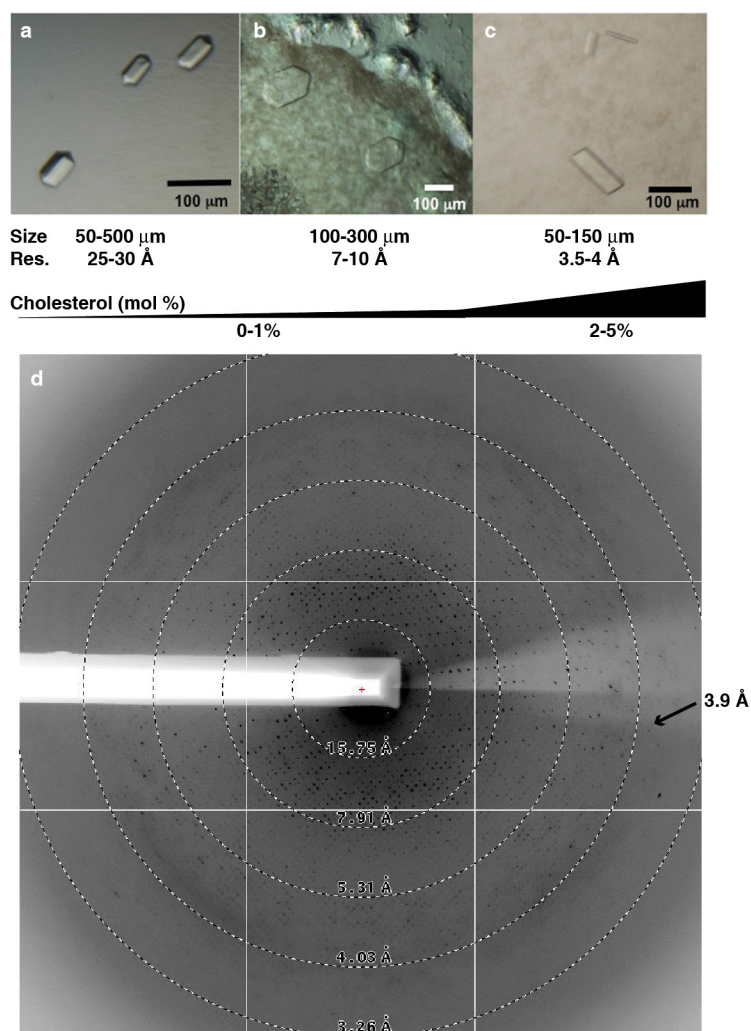
- Chloupková, M. *et al.* Expression of 25 human ABC transporters in the yeast *Pichia pastoris* and characterization of the purified ABCG3 ATPase activity. *Biochemistry* **46**, 7992–8003 (2007).
- Zhang, D.-W., Graf, G. A., Gerard, R. D., Cohen, J. C. & Hobbs, H. H. Functional asymmetry of nucleotide-binding domains in ABCG5 and ABCG8. *J. Biol. Chem.* **281**, 4507–4516 (2006).
- Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data. *Methods Enzymol.* **276**, 307–326 (1997).
- Minor, W., Cymborowski, M., Otwinowski, Z. & Chruszcz, M. HKL-3000: the integration of data reduction and structure solution – from diffraction images to an initial model in minutes. *Acta Crystallogr. D* **62**, 859–866 (2006).
- Otwinowski, Z., Borek, D., Majewski, W. & Minor, W. Multiparametric scaling of diffraction intensities. *Acta Crystallogr. A* **59**, 228–234 (2003).
- Borek, D., Minor, W. & Otwinowski, Z. Measurement errors and their consequences in protein crystallography. *Acta Crystallogr. D* **59**, 2031–2038 (2003).
- Borek, D., Ginell, S. L., Cymborowski, M., Minor, W. & Otwinowski, Z. The many faces of radiation-induced changes. *J. Synchrotron Radiat.* **14**, 24–33 (2007).
- Borek, D., Cymborowski, M., Machius, M., Minor, W. & Otwinowski, Z. Diffraction data analysis in the presence of radiation damage. *Acta Crystallogr. D* **66**, 426–436 (2010).
- Borek, D., Dauter, Z. & Otwinowski, Z. Identification of patterns in diffraction intensities affected by radiation exposure. *J. Synchrotron Radiat.* **20**, 37–48 (2013).
- Sheldrick, G. M. Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. *Acta Crystallogr. D* **66**, 479–485 (2010).
- Otwinowski, Z. Maximum likelihood refinement of heavy atom parameters. In *Proc. CCP4 Study Weekend: Isomorphous Replacement and Anomalous Scattering* (Eds Evans, P. R., Wolf, W., Leslie, A. G. W.) 80–86 (Daresbury Laboratory: Science and Engineering Research Council, 1991).
- Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* **67**, 235–242 (2011).
- Cowan, K. & Main, P. Miscellaneous algorithms for density modification. *Acta Crystallogr. D* **54**, 487–493 (1998).
- Zhang, K. Y., Cowtan, K. & Main, P. Combining constraints for electron-density modification. *Methods Enzymol.* **277**, 53–64 (1997).
- Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
- Cowan, K. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr. D* **62**, 1002–1011 (2006).
- Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).

48. Nicholls, R. A., Fischer, M., McNicholas, S. & Murshudov, G. N. Conformation-independent structural comparison of macromolecules with ProSMART. *Acta Crystallogr. D* **70**, 2487–2499 (2014).
49. Davis, I. W. *et al.* MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35**, W375–W383 (2007).
50. Chivian, D. *et al.* Prediction of CASP6 structures using automated Robetta protocols. *Proteins* **61** (Suppl. 7), 157–166 (2005).
51. Stone, D. K., Xie, X. S. & Racker, E. Inhibition of clathrin-coated vesicle acidification by duramycin. *J. Biol. Chem.* **259**, 2701–2703 (1984).
52. Wang, J. *et al.* Relative roles of ABCG5/ABCG8 in liver and intestine. *J. Lipid Res.* **56**, 319–330 (2015).
53. Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl Acad. Sci. USA* **110**, 15674–15679 (2013).
54. Ovchinnikov, S. *et al.* Large-scale determination of previously unsolved protein structures using evolutionary information. *eLife* **4**, e09248 (2015).
55. Vehlou, C. *et al.* CMView: interactive contact map visualization and analysis. *Bioinformatics* **27**, 1573–1574 (2011).
56. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
57. Bayly, C. I., Cieplak, P., Cornell, W. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **97**, 10269–10280 (1993).
58. Case, D. A. *et al.* The Amber biomolecular simulation programs. *J. Comput. Chem.* **26**, 1668–1688 (2005).
59. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **25**, 247–260 (2006).
60. Wickstrom, L., Okur, A. & Simmerling, C. Evaluating the performance of the ff99SB force field based on NMR scalar coupling data. *Biophys. J.* **97**, 853–856 (2009).
61. Dickson, C. J. *et al.* Lipid14: the amber lipid force field. *J. Chem. Theory Comput.* **10**, 865–879 (2014).
62. Sagui, C., Pedersen, L. G. & Darden, T. A. Towards an accurate representation of electrostatics in classical force fields: efficient implementation of multipolar interactions in biomolecular simulations. *J. Chem. Phys.* **120**, 73–87 (2004).
63. Ryckaert, J. P., Ciccotti, G. & Berendsen, H. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. *J. Comput. Phys.* **23**, 321–341 (1977).
64. Uberuaga, B. P., Anghel, M. & Voter, A. F. Synchronization of trajectories in canonical molecular-dynamics simulations: observation, explanation, and exploitation. *J. Chem. Phys.* **120**, 6363–6374 (2004).
65. Bahar, I., Lezon, T. R., Bakan, A. & Shrivastava, I. H. Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. *Chem. Rev.* **110**, 1463–1497 (2010).
66. Pei, J. & Grishin, N. V. PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information. *Methods Mol. Biol.* **1079**, 263–271 (2014).
67. Katoh, K. & Standley, D. M. MAFFT: iterative refinement and additional methods. *Methods Mol. Biol.* **1079**, 131–146 (2014).
68. Pei, J. & Grishin, N. V. AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics* **17**, 700–712 (2001).
69. Lazarevic, V. & Karamata, D. The tagGH operon of *Bacillus subtilis* 168 encodes a two-component ABC transporter involved in the metabolism of two wall teichoic acids. *Mol. Microbiol.* **16**, 345–355 (1995).
70. Evans, I. J. & Downie, J. A. The *nodI* gene product of *Rhizobium leguminosarum* is closely related to ATP-binding bacterial transport proteins; nucleotide sequence analysis of the *nodI* and *nodJ* genes. *Gene* **43**, 95–101 (1986).
71. Goldman, B. S., Beck, D. L., Monika, E. M. & Kranz, R. G. Transmembrane heme delivery systems. *Proc. Natl Acad. Sci. USA* **95**, 5003–5008 (1998).
72. Ajdic, D. *et al.* Genome sequence of *Streptococcus mutans* UA159, a cariogenic dental pathogen. *Proc. Natl Acad. Sci. USA* **99**, 14434–14439 (2002).
73. Han, C. S. *et al.* Pathogenomic sequence analysis of *Bacillus cereus* and *Bacillus thuringiensis* isolates closely related to *Bacillus anthracis*. *J. Bacteriol.* **188**, 3382–3390 (2006).



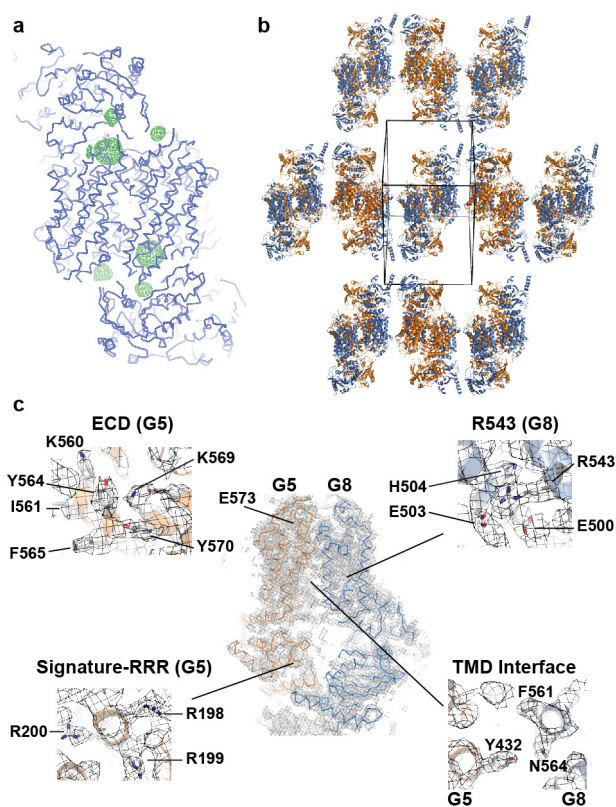
Extended Data Figure 1 | Purification and ATPase activity of G5G8 heterodimers.

a, G5G8 heterodimers were co-purified by sequential affinity chromatography (Ni-NTA and CBP) followed by gel-filtration chromatography using Superdex 200 (SD 200) as described in the Methods. Protein purity was analysed by sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS–PAGE) and Coomassie staining. **b**, Dimerization of detergent-soluble G5G8 was analysed by size fractionation using blue native PAGE (BN–PAGE) followed by Coomassie staining of the gel (top). Proteins were transferred to nitrocellulose and immunoblotting was performed using an anti-RGSH₄ monoclonal antibody (Qiagen) and anti-G8 rabbit polyclonal antibodies (bottom). Arrows point to the G5G8 dimer. **c**, Bile acid stimulated ATPase activity of purified recombinant WT G5 and G8 (4 μg) and WT G5 and catalytic-deficient G8 (G8-G216D) (Mut, 8 μg). Each point represents the mean ± s.d. of three separate measurements.

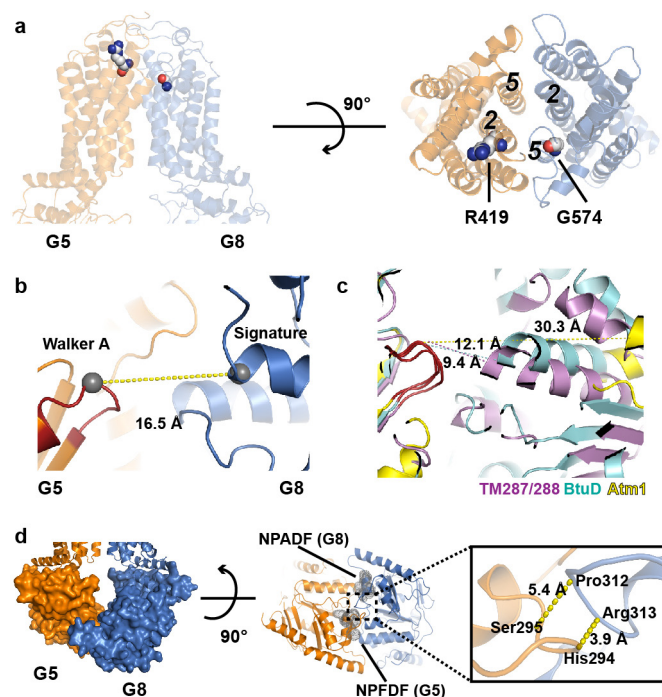


Extended Data Figure 2 | Optimization of crystal growth and X-ray diffraction. Crystals of G5G8 were grown using bicelle crystallization (see Methods). Optimal diffraction-quality crystals were obtained by using cholesterol as additive to the bicelle matrices. Shown here are crystals observed under a light microscope without the polarizer. Scale

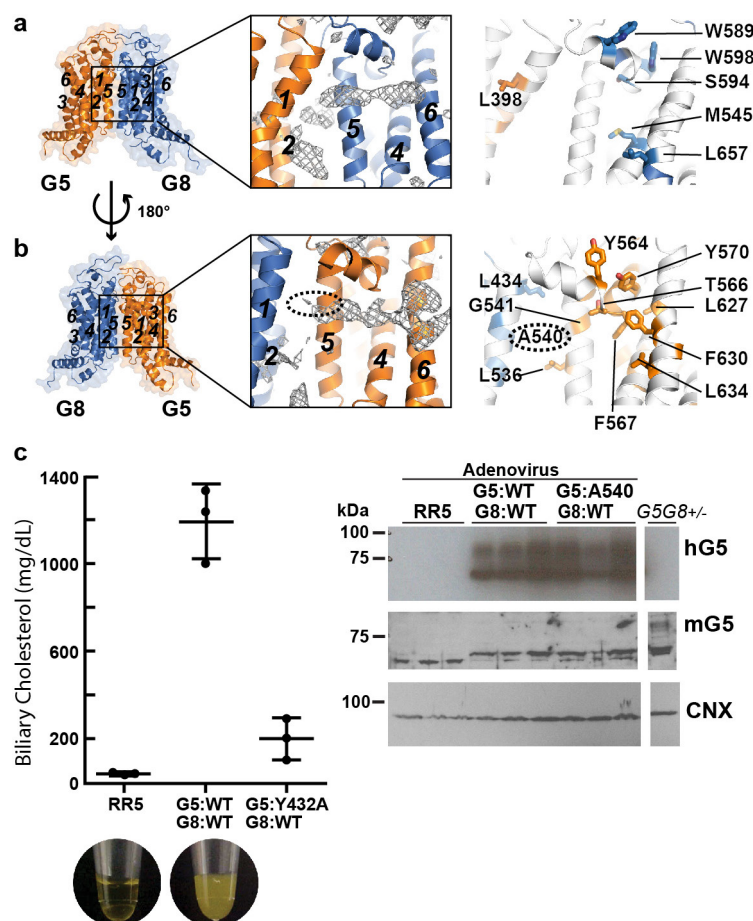
bar, 100 μm. **a, b**, In the presence of 0–1% (mol) cholesterol, crystals grew up to 500 μm, but only diffracted to 7–10 Å. **c**, In the presence of 2–5% (mol) cholesterol, crystals grew to 50–150 μm (longest edge) and diffracted to 3.5–4 Å. **d**, A diffraction frame shows structural information of G5G8 crystals out beyond 3.9 Å resolution.



Extended Data Figure 3 | Experimental electron density maps and crystal packing in G5G8 bicelle crystals. **a**, Side view of C α backbone for two G5G8 dimers in an asymmetric unit (containing two transporters) with the anomalous Fourier map calculated from W-clusters plotted at 3 σ . **b**, Orthorhombic G5G8 bicelle crystals with lateral crystal packing of TMDs indicative of G5G8 reconstitution into phospholipid bilayers (G5: orange; G8: blue). **c**, Electron density maps ($2F_o - F_c$) show the complete G5G8 heterodimer (centre, contoured at 1.5 σ), selected residues at the G5 ECD (upper left, 1.5 σ), conserved three arginines of G5 (R198/R199/R200) in proximity to the Signature motif (bottom left, 1.0 σ), R543 of G8 where point mutation R543S causes sitosterolaemia (upper right, 1.5 σ), and Y432 on TMH5 (G5) and F561 on TMH5 (G8) at the TMD interface (bottom right, 1.5 σ).

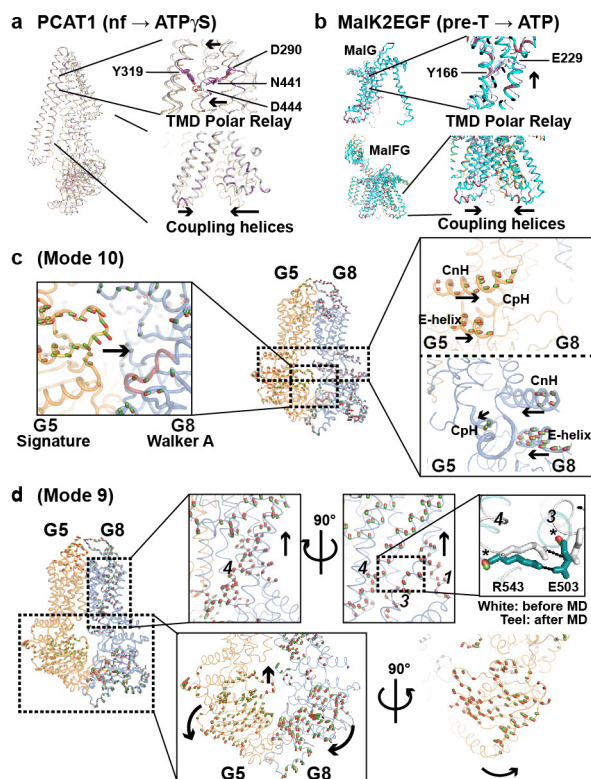


Extended Data Figure 4 | G5G8 heterodimer interface. **a**, Sitosterolaemia mutation positions R419 (G5) and G574 (G8) at the G5-TMH2 and G8-TMH5 interface near the extracellular surface. **b**, Separation between the +4 glycine of G5 Walker A (G5, GSSGSGKT) and G8 Signature (LSGGE) motifs. **c**, Superposition of the nucleotide-free NBDs of TM287/288 (PDB accession number 4Q4H), BtuD (PDB accession number 2QI9), and Atm1 (PDB accession number 4MYC) with varying distances between NBDs. **d**, Left: surface-filled view of the NBDs of G5G8 showing contacts between the two half-transporters at the cytoplasmic apex. Middle: interaction between conserved NPFDF (G5) and NPADF (G8) sequences (grey dots). Right: closest distances between the two half-transporters.

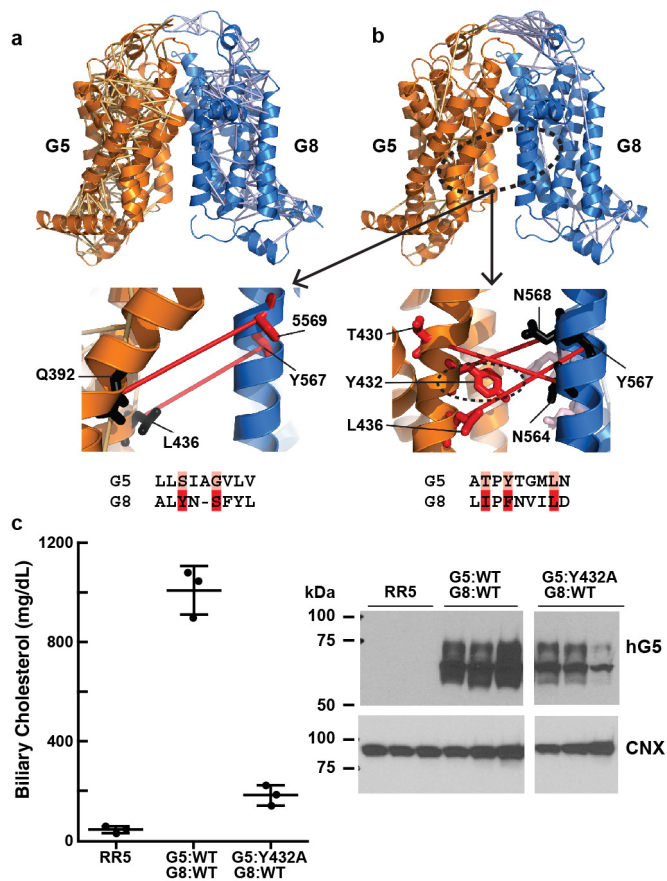


Extended Data Figure 5 | Vestibules in the membrane spanning region. **a, b**, Left: cartoon of G5G8 TMDs in transparent surface (orange and blue). Vestibules on opposing faces of the TMDs are highlighted in boxes. Middle: $F_o - F_c$ difference electron density map contoured at 3.0σ showing extended features at the vestibules. TMHs are numbered. Right: highly conserved residues in G5 and G8 (sequence conservation ≥ 7 , see Methods) are shown as orange (G5) and blue (G8) cartoon/sticks. **c**, Left: *in vivo* functional reconstitution assay using G5G8 KO mice (G5G8^{-/-} mice)¹⁹. Biliary cholesterol levels from mice infected with RR

(empty adenovirus), or with adenoviruses expressing human (h)G5:WT and hG8:WT, or with adenoviruses expressing hG5:A540F and hG8:WT are shown. Pictures of the bile are below the graph. Each experiment represents the mean \pm s.d. from three infected mice ($n=3$) in each group. Right: expression of hG5 and hG8 detected by immunoblotting using anti-hG5 monoclonal antibodies (see Methods). Connexin was used as the gel-loading control. As a positive control for the mouse and a negative control for the human anti-G5 antibody, we also immunoblotted liver membranes from G5G8^{+/-} mice.

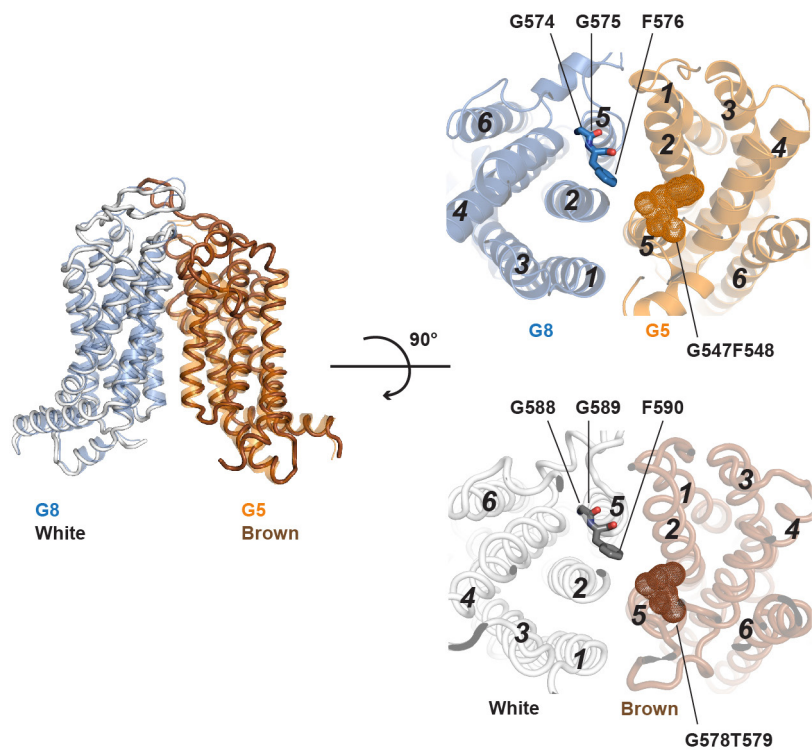


Extended Data Figure 6 | TMD polar relay in other ABC transporters and molecular dynamics simulation of G5G8 structure. **a, b,** Analysis of bacterial PCAT1 (**a**) and maltose transporter (MalK₂EGF, **b**) structures reveals conserved polar residues in the TMD subunits. TMD superposition of a nucleotide-free structure with a nucleotide-bound structure shows breakage of hydrogen bonding in TMD polar relay (top right) and inward movement of coupling helices (bottom right). (Nucleotide-free PCAT1 (PDB accession number 4RY2), ATP- γ S-bound PCAT1 (PDB accession number 4s0f), nucleotide-free MalK₂EGF (PDB accession number 3PV0), and ATP-bound MalK₂EGF (PDB accession number 2R6G).) **c, d,** Molecular dynamics simulation (100 ns) and vibrational mode analysis performed on the nucleotide-free G5G8 structure (see Methods). For each vibrational mode, the contribution of each residue was calculated by adding up the vectors of the main chain atoms. Among the lowest 20 vibrational modes, modes 9 and 10 are shown. Mode 10 (**c**) describes collective movements (green-to-red vectors on the C α atoms) between CpH/CnH/E-helix bundle and the interface at ATP-binding cassette, indicating inward movement of the CpH/CnH/E-helix bundles from both G5 and G8. This is consistent with stereotypical inward motion of coupling helices in bacterial ABC transporters. Mode 9 (**d**) describes a collective movement (arrows) of NBD (bottom) and TMH1, 3, and 4 in G8 (top), indicating coupling between NBD and TMD. R543 of G8, a conserved polar residue in the TMD polar relay where a disease-causing mutation occurs, is highlighted together with its interacting partner, E503.



Extended Data Figure 7 | Coevolution analysis of ABCG5 and ABCG8.

a, Co-evolving residue pairs from the ABCG5 and ABCG8 TMD that were closer than an all-atom cutoff of 8 Å (≤ 8 Å) are depicted as lines connecting C α atoms of the residues in contact in the respective TMD structures. **b**, Co-evolving pairs that are more spatially distant than the defined cutoff (>8 Å). Expanded views show potential interface contacts predicted by spatially distant G5 and G8 coevolving residue pairs (red lines connecting C α atoms) (top). Cross contacts in G8 (red) were mapped from the distant G5 GREMLIN pairs near the interface (pink) using the GREMLIN MSA, and vice versa (bottom). For example, we mapped S538 from the predicted G5 pair L436/S538 to the corresponding residue in G8-Y567, and considered G5-L436 and G8-Y567 to be a potential interface contact. Predicted interface contacts that are distant in the nucleotide-free structure could form contacts at another stage of the ABCG ATPase cycle. **c**, Left: *in vivo* functional reconstitution assay using liver-specific G5G8^{-/-} mice⁵². Biliary cholesterol levels of mice infected with RR (empty adenovirus), or with adenoviruses expressing hG5:WT and hG8:WT, or with adenoviruses expressing hG5:Y432A and hG8:WT are shown. G5-Y432 is part of the TMD polar relay and predicted to interact with G8-N568 during the sterol transport cycle (**b**). Each experiment represents the mean \pm s.d. from three infected mice ($n = 3$) in each group. Right: immunoblot analysis of hG5 levels in membranes isolated from the mice using an anti-hG5 monoclonal antibody (see Methods). Connexin was used as the gel-loading control.



		TMH5	
G5	(Homo sapiens)	527	PNIVNSVVALLSIAGVLVGSGLRN 551
G5	(Danio rerio)	531	PNMVNSGVALLNIAGIMVGSGLRG 555
G8	(Homo sapiens)	556	FHMASFFSNALYN-SFYLAGGFMIN 579
G8	(Danio rerio)	537	LQTSSFMGNALFT-VFYLTAGFVIS 560
White		570	TSMALSVGPPVII-PFLLF GGF FLN 593
Brown		559	DKMASECAAPFDL-IFLI FG GTYMN 582
Scarlet		550	VPLAMAYLVPLDY-IFMITSGIFIQ 573
			*

Extended Data Figure 8 | Homology model of *Drosophila* white/brown based on the G5G8 structure. A model of the *Drosophila melanogaster* white/brown heterodimeric guanine pigment transporter was made using the program Modeller²⁸. On the basis of the sequence alignment and comparison of scoring metrics for the two possible heterodimer models (that is, white modelled on G5 versus G8, brown on the other), we selected

G8 (blue cartoon) as the white template and G5 (orange cartoon) as the brown template. At right, top/extracellular views of the G5G8 structure and white/brown model are shown, with residues important for dye interaction highlighted. At bottom is an MSA for the C-terminal region of TMH5, showing the conservation of the functionally important amino acids.

Extended Data Table 1 | Data processing and refinement statistics

	Native*	[PW ₁₂ O ₄₀ ³⁻] [†]	[(CH ₃) ₃ Pb ⁻]	(Ta ₆ Br ₁₂ ²⁺)
Data collection				
Beamline	19-ID-D/23-ID-D [‡]	19-ID-D	19-ID-D	19-ID-D
Space group	I 222	I 222	I 222	I 222
Cell dimensions				
<i>a</i> , <i>b</i> , <i>c</i> (Å)	173.6, 224.8, 253.3	175.5, 227.5, 254.5	174.6, 225.9, 253.4 [§] 173.6, 225.9, 252.7	176.0, 228.0, 253.7
Resolution (Å)	50-3.9 (3.93-3.9)	50-5.0 (5.04-5.0)	50-4.5 (4.54-4.5)	50-5.0 (5.04-5.0)
<i>R</i> _{sym} or <i>R</i> _{merge}	16.1 (NA)	13.5 (33.5)	8.7 (NA) [§] 7.1 (NA)	8.8 (NA)
<i><I>/<σI></i>	8.8 (0.15)	5.1 (1.4)	8.0 (0.45) [§] 6.1 (0.18)	8.6 (0.50)
Completeness (%)	99.4 (84.2)	94.4 (47.3)	97.4 (55.9) [§] 94.7 (54.7)	81.1 (18.3)
Redundancy	18.9 (2.5)	3.1 (1.7)	6.0 (2.3) [§] 4.3 (2.5)	3.7 (1.3)
Refinement				
Resolution (Å)	25-3.94			
No. reflections	34889			
<i>R</i> _{work} / <i>R</i> _{free}	24.5 / 32.9			
No. atoms				
Protein	18151			
R.m.s deviations				
Bond lengths (Å)	0.010			
Bond angles (°)	1.64			

*Number of crystals: 19.

†Number of crystals: 2.

‡MAR CCD detector was used.

§Data collected at peak wavelength: 13.053 keV.

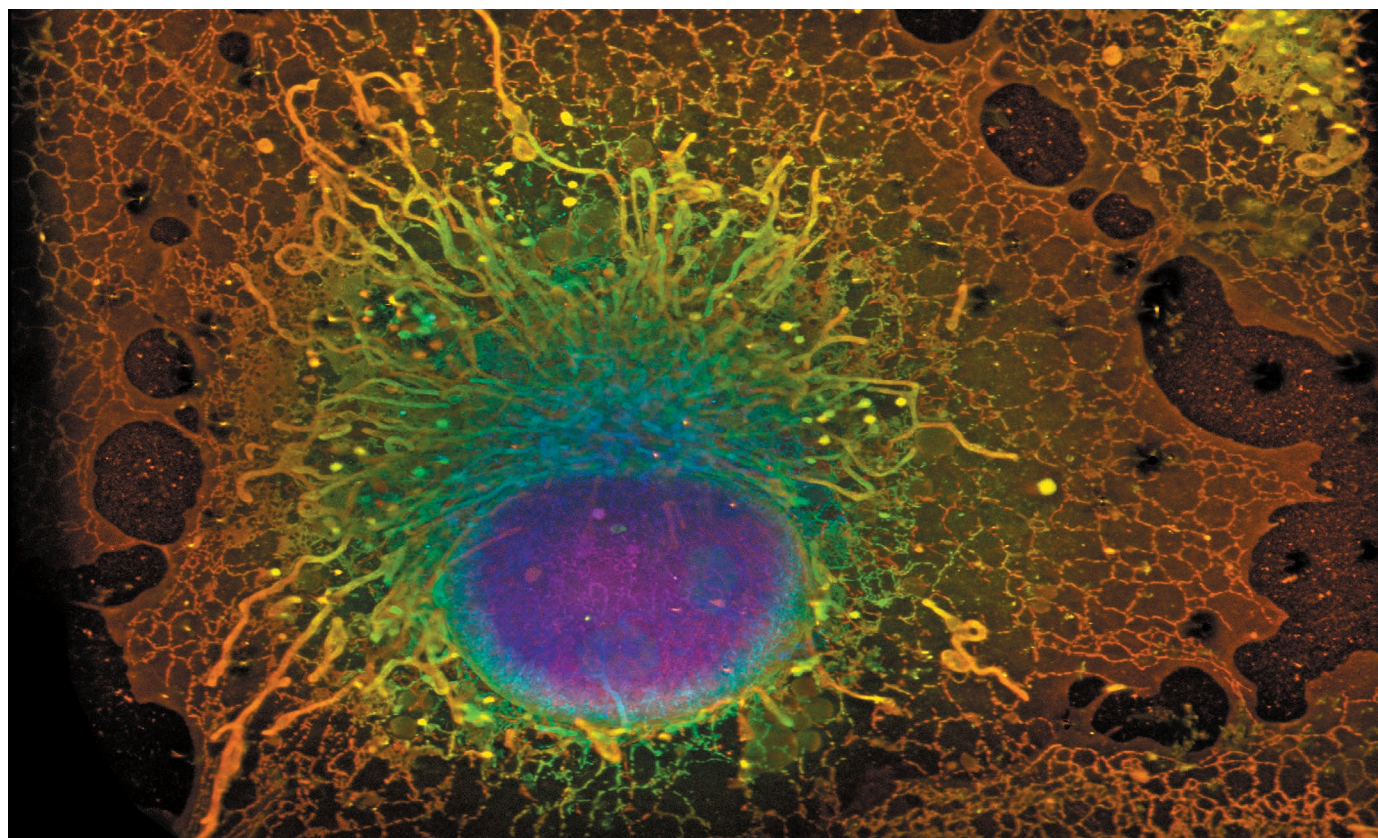
||Data collected at inflection wavelength: 13.043 keV.

TECHNOLOGY FEATURE

ILLUMINATING LIFE'S BUILDING BLOCKS

A suite of tools now enables scientists to see proteins at work in living cells at the single-molecule level.

WESLEY R. LEGANT



A combination of fluorescent labelling and light-sheet imaging yields a super-resolution view of structures inside a fixed (non-living) cell.

BY MARISSA FESSENDEN

Biophysicist Joerg Bewersdorf says that 2006 was fluorescence microscopy's *annus mirabilis* — a 'miraculous year' as momentous in its own way as 1905, when Albert Einstein revolutionized physics in the realms of relativity, quantum theory and atomic physics. In microscopy's case, the revolution consisted of three papers^{1–3} that, for the first time, gave scientists the power to peer down into the cell and track the behaviour of individual molecules.

"Every molecule is a machine, a little nano-machine," Bewersdorf says. Proteins, in particular, are complex molecules that twist, flex,

open and shut in a multitude of ways to perform the reactions necessary for cell metabolism and growth, sending messages and providing structure. "That is what we are ultimately interested in understanding," says Bewersdorf: "How do all these little machines work together for the global function of the cell?"

Until scientists could observe that world, however, they had only the cloudiest idea of how to answer that question. Light microscopes were no help; beyond a certain magnification, diffraction causes light waves to spread out instead of converging to form an image. Any features closer together than about 200 nanometres, or about 40 times the width of a typical cell membrane, become a hopeless blur. Images

made using electron microscopy can resolve fine structures — but they are static and almost impossible to obtain from a live cell.

The three laboratories that independently circumvented the 'diffraction barrier' in 2006 adopted a similar strategy: studying the sample with specialized fluorescence probes that can be selectively switched on, a few at a time, until all of the probes are captured in a series of images. Combining the data from those images builds a picture, in a similar way to an impressionist painter building up a scene with dots of colour (see 'Connecting the dots'). The three techniques — photoactivated localization microscopy (PALM)¹, fluorescence PALM (FPALM)² and stochastic optical

► reconstruction microscopy (STORM)³ — can differentiate between points just 20 nanometres apart, producing the sharpest-ever fluorescent images at the single-molecule level. Researchers have rushed to take advantage of these capabilities. At Bewersdorf's lab at Yale University in New Haven, Connecticut, for example, he has filmed proteins moving across the surface of living cells⁴.

In the decade since 2006, the three techniques have inspired a wave of technological and methodological innovations. Researchers are designing fluorescent probes that shine brighter and are robust enough to image cellular processes as they unfold. They are also developing methods that cause less disruption to living cells. Several illumination strategies seek to reduce the visual noise caused by background fluorescence, whereas computational methods and strategies are allowing researchers to combine multiple imaging approaches to see molecular interactions in real time.

"The big excitement over these past few years is that these technologies have become doable in living cells," says Jennifer Lippincott-Schwartz, a cell biologist at the Howard Hughes Medical Institute Janelia Research Campus in Ashburn, Virginia. "The time is definitely ripe for being able to image individual proteins using fluorescence."

LASTING LONGER, SHINING STRONGER

All three of the super-resolution microscopy techniques invented in 2006 rely on the light emitted by probe compounds such as green fluorescent protein (GFP), which was first isolated in a bioluminescent jellyfish. The genes for these probes can be inserted into the DNA coding for a cellular protein of interest. Then, when that protein is produced, it will have the fluorescent compound attached, and will reveal its presence by glowing.

But these techniques have some severe limitations. A big one is that many of these probes can emit only a finite number of photons before they are irreversibly damaged by the intensity of the lasers that are used to excite

them into emitting light. Even before this photo-bleaching effect takes hold, the probes are quite dim when imaged individually.

Synthetic versions of these probes known as organic (that is, carbon-containing) dyes are brighter, but they cannot be genetically encoded to their target and manufactured inside the cell. Instead, they are often linked with antibodies that can seek out the protein of interest. However, that combination can make the probes too large to pass through the cell membrane or bulky enough to interfere with a protein's function.

"Probes are really limiting us and really defining to some extent where this technology can go," says Bewersdorf.

"The time is definitely ripe for being able to image individual proteins using fluorescence."

Fortunately, alternatives are emerging. Bewersdorf's group is working with two 'clickable chemistry' probes: SNAP-tag, from New England

Biolabs in Ipswich, Massachusetts, and HaloTag, from Promega in Madison, Wisconsin. These technologies involve a short target sequence that can be encoded into the protein of interest and a dye molecule that clicks into place with its target protein through a simple chemical reaction. Bewersdorf and his colleagues demonstrated that the two tags can be used with organic dyes in living cells to achieve a resolution below 50 nm⁵ — almost as precise as the 20-nm resolution of the original techniques, with the advantage that they combine the specificity and leanness of genetically encoded probes with the brightness of synthetic dyes.

Researchers have also turned to quantum dots, nanoscale semiconductors that are not only bright and stable for a month or more, but can also link to biological molecules. Diane Lidke, a biophysicist at the University of New Mexico in Albuquerque, uses quantum dots in her lab's work on cell signalling. For her, the benefits outweigh the major disadvantage of quantum dots — their size. Commercially available quantum dots are surrounded by a shell that can link the dot to other

molecules but that expands the dot's diameter to 15–25 nm. "They are quite large and bulky," she says, at least when compared with a fluorescent protein, which can be just 4 nm wide.

That pitfall means that researchers have difficulty getting quantum dots into the cell or other tight spaces, but they work well for the kind of extracellular, membrane-bound proteins that Lidke targets. In collaboration with her husband Keith Lidke, a physicist at the same university, she has developed a multiple-colour, fast, single-molecule tracking method that uses quantum dots to produce images on a custom microscope⁶.

Still, much of the cell's internal processes remain locked inside the membrane, difficult to reach with fluorescent probes.

CRACKING OPEN THE CELL

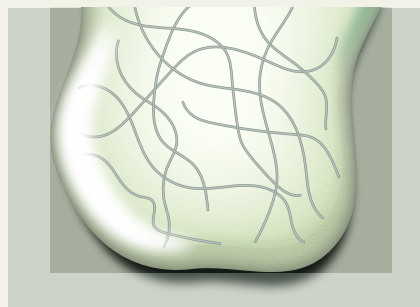
Getting past the cell's membrane is one of the most daunting hurdles that fluorescence microscopy faces. "Even though it is only five nanometres thick, [the cell membrane] has had a few billion years of evolution to separate the inside of the cell from the outside, and it does this amazingly well," says Paul Selvin, a biophysicist at the University of Illinois at Urbana-Champaign.

Selvin's lab has developed its own version of quantum dots that are smaller, closer to 9 nm in diameter⁷. That size reduction helps him to slip quantum dots into the approximately 20–40 nm gap between nerve cells, the synapse, where signalling molecules pass on messages to nearby nerve cells. Once in that cleft, the quantum dots can bind to and advertise the presence of receptors that facilitate memory formation. Selvin hasn't sent these smaller quantum dots inside living cells yet, but he says that this could be possible.

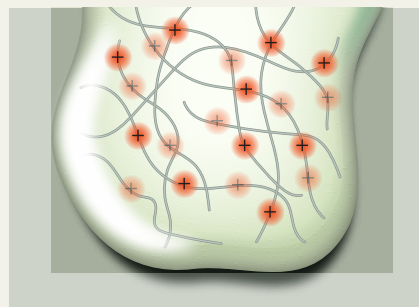
Selvin's lab is also working on a strategy to punch holes in the plasma membrane and then quickly reseal them so that the cell isn't disturbed. "We effectively drill little microscopic pores into the membrane that are about five nanometres," he says, using a bacterial enzyme called streptolysin O. That's just wide enough

CONNECTING THE DOTS

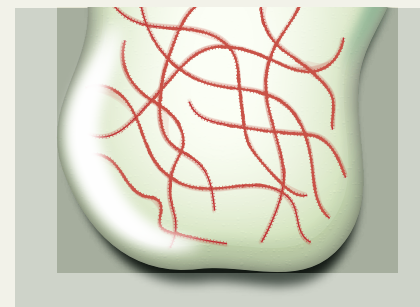
Super-resolution microscopy based on fluorescent probe molecules can reveal cellular structures that are much smaller than the wavelength of light.



The cellular structures of interest, shown here as grey strands, are labelled by fluorescent probes that are invisible until they are illuminated.



Low-intensity laser light repeatedly excites individual probes at random, causing them to emit isolated blobs of fluorescent light (red).



Each probe molecule is mapped to the precise centre of its blob. Over time, these points build up a map of the original structures.

to let a fluorescent protein, even one linked to an antibody, slip inside and find its intracellular target. The method, which Selvin has yet to publish, then patches up those holes within 20 minutes.

Yet there is always the concern that the added probe could interfere with the target protein's typical function. An alternative strategy that doesn't impair the protein comes from Jie Xiao, a biophysicist at Johns Hopkins University in Baltimore, Maryland. Her probe molecules are genetically encoded, but instead of hanging on to the molecule of interest, they are cleaved by an enzyme as soon as they are produced and scurry off to a particular part of the cell membrane. That means that they no longer carry any information about the target molecule's position, but they are in a position where Xiao can count them precisely and thus get an exact tally of the proteins produced, while the proteins themselves are free to go about their business unencumbered. She calls the method co-translational activation by cleavage (CoTrAC)⁸.

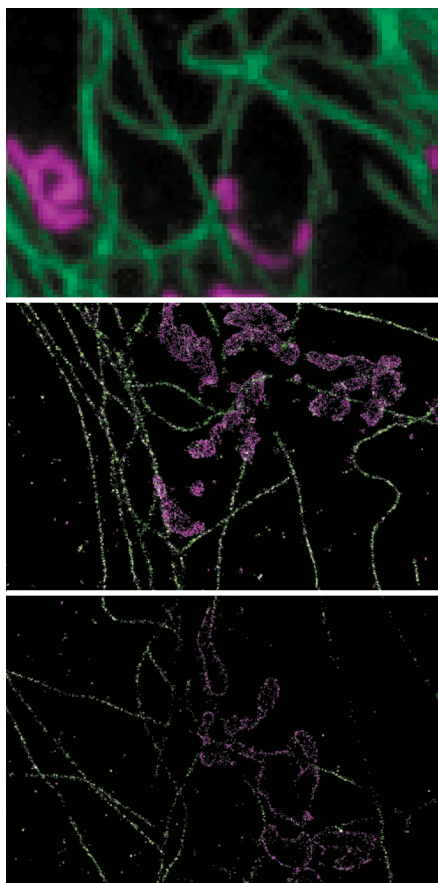
"Being able to quantify the absolute protein level in a living cell is very important," Xiao explains. "Most of the time people use the fluorescence to indicate relative change." However, only a few of the gene-regulation proteins she studies are produced at a time, which makes them difficult to image with most super-resolution techniques. Furthermore, small changes in the exact number of those proteins can determine whether the cell changes states or not. The advantage of CoTrAC is that, by gathering probes for different proteins at different locations on the cell membrane, this technique can be used to count multiple protein products using the same fluorescence molecule and colour. This is a crucial ability when producing different probes can take varying amounts of time and may obscure the timing of cell processes.

LIGHTING THE WAY

Crisper, striking images can come from brighter probes, but another way to make the probes stand out is to reduce the background light. "You have many different planes that you observe simultaneously but you can see only one plane sharply — that in the focus plane of your camera," says Ulrich Kubitschek, a biophysical chemist at the University of Bonn in Germany. "But you still have the diffused background of everything else in the cell."

Non-target proteins can even have their own, natural, dim fluorescence that contributes to this noise. If that background can instead be kept in the dark, the contrast and clarity of images can be enhanced.

Therefore, researchers are constantly improving their illumination strategies. Kubitschek's lab uses light-sheet microscopy to generate a very thin beam of precisely focused light that slices through the sample from the side. "We have a glass chamber that is transparent from below and the sides," he says.



Fluorescently labelled microtubules (green) and mitochondria (magenta) are blurred when seen through a conventional microscope (top). They are sharper when stochastic optical reconstruction microscopy (STORM, middle) is used — and sharper still (bottom) when an advanced version of STORM reveals structures in 3D.

By shining light through the side rather than the top of the sample, his group illuminates a section only 200–300 nm thick and observes the sample from below. In this way, the group has seen RNA molecules as they are exported out of the nucleus through a protein complex called the nuclear pore and into the cytoplasm, where they go on to instruct protein synthesis⁹.

Microscopes equipped to carry out light-sheet imaging are commercially available from companies such as Leica Microsystems in Wetzlar, Germany, and Carl Zeiss in Oberkochen, Germany. Groups with the necessary expertise can even do as Kubitschek's team has done and build their own, custom microscopes.

The next iteration of such selective plane illumination microscopy is called lattice light-sheet microscopy and hails from the lab of Eric Betzig, a physicist at Janelia who also developed PALM microscopy. The technology can generate an illumination plane 300–500 nm thick, says collaborator Zhe Liu, a cell biologist at Janelia — but the real advantage of the method is the structure that the light takes. The lattice forms a three-dimensional grid that moves, illuminating successive sections through the sample.

“You can image for a much longer period of time because the [out-of-focus] molecules are being preserved,” Liu says. Capturing 3D images is also possible. Liu first started using the technology three years ago, while Betzig was still developing it, to examine the organization of protein clusters that are necessary for maintaining a stem cell’s ability to self-renew¹⁰.

Lattice light-sheet microscopy is not yet commercially available, although Carl Zeiss is working on a microscope. Until this goes on sale, groups that are interested in using the technology need to assemble their own custom microscopes. “Alignment is tricky,” Liu cautions, but Betzig and others provide workshops to help those willing to tackle the challenge.

MOTION IN MINIATURE

These probes and illumination strategies can be combined for truly novel insights. “Typically, the field has defined the structures in cells,” says Lippincott-Schwartz. “Now we are getting a handle on the underlying mechanisms that allow these structures to move and interact.”

In her lab at Janelia, Lippincott-Schwartz and her colleagues are leveraging lattice light-sheet microscopy — which she calls a “truly transformative technology” — to look at the way cell organelles and proteins interact. She and her colleagues have watched enzymes repeatedly interact with the endoplasmic

reticulum (ER), a network of membranes in the cell where proteins are synthesized and folded¹¹. Canonically, the ER has been thought of as only a site of protein secretion, but these observations have “really made me start thinking very differently about the primary or major function of this organelle”, she says. The ER might be communicating with other structures more than was previously thought.

This kind of work requires expertise even beyond that needed to align and use the microscopes properly. Lippincott-Schwartz explains that it requires the appropriate algorithms to reconstruct the image from the data acquired by the microscopes. “This is not just something you can pick up,” she says. Typically, labs without substantial experience in super-resolution imaging and the statistical methods it relies on can turn to the specialized knowledge held by experts in the shared imaging facilities that some institutions have. But she also sees a need for standards related to image acquisition and analysis to be set for the field. “Otherwise, there will be information that will be improperly interpreted,” she says.

The complexity of single-molecule imaging means that clean data, controls and proper analysis are invaluable. “If I could recommend something to people entering the field,” says Antoine Triller, a neurobiologist at the Institute of Biology of the École Normale Supérieure in

Paris who studies the movement of molecules in the synapse, “it would be either to have a good background in statistical physics or to work with people with a very good knowledge in the field.”

Making this effort is worth it, however, to gain access to the ‘black box’ of molecular-scale life occupying the space between classic light microscopy at the microscale and electron microscopy at the nanoscale. Single-molecule approaches in living cells offer a way to find new biological parameters to measure and observe, Triller says. “These parameters will allow us to develop new theoretical fields and a new understanding of living matter.” ■

Marissa Fessenden is a freelance writer in Bozeman, Montana.

1. Betzig, E. *et al.* *Science* **313**, 1642–1645 (2006).
2. Hess, S. T., Girirajan, T. P. K. & Mason, M. D. *Biophys. J.* **91**, 4258–4272 (2006).
3. Rust, M. J., Bates, M. & Zhuang, X. *Nature Methods* **3**, 793–796 (2006).
4. Juetten, M. F. & Bewersdorff, J. *Nano Lett.* **10**, 4657–4663 (2010).
5. Bottanelli, F. *et al.* *Nature Commun.* **7**, 10778 (2016).
6. Cutler, P. J. *et al.* *PLoS ONE* **8**, e64320 (2013).
7. Ma, L. *et al.* *J. Am. Chem. Soc.* **138**, 3382–3394 (2016).
8. Hensel, Z., Fang, X. & Xiao, J. J. *Vis. Exp.* **73**, e50042 (2013).
9. Spille, J.-H. *et al.* *Nucleic Acids Res.* **43**, e14 (2015).
10. Liu, Z. *et al.* *eLife* **3**, e04236 (2014).
11. Sengupta, P. *et al.* *Proc. Natl Acad. Sci. USA* **112**, E6752–E6761 (2015).

CAREERS

AFTER THE PHD How PhD graduates move through and plan careers **p.570**

MOVING INTO INDUSTRY A medicinal chemist shares his story go.nature.com/dytig2

NATUREJOBS For the latest career listings and advice www.naturejobs.com



CHRIS RYAN/GETTY

For PhD holders who wish to move beyond the laboratory, an MBA might be essential.

EDUCATION

Degrees of success

An MBA can unlock progress to the higher ranks of a company — and many firms are willing to pay for one.

BY CHRIS WOOLSTON

Life-science PhD graduates who wish to leave academia often find rewarding careers in the laboratories of biotechnology and pharmaceutical companies. But some find that the lab isn't enough. Researchers who choose to move beyond the bench to the upper levels of the company often decide to add three more letters to their CV: MBA.

Investing time and money in another degree may seem an unappealing prospect for many PhD holders, but that's the reality of the competitive job market: sometimes you have to go beyond the usual training to get the job. An MBA (master of business administration) can open up career possibilities for a biotechnology or drug-development researcher and help them to stand out from the crowd. Those who decide to take the plunge face key questions:

how and when to pursue an MBA (see 'When to go for an MBA'), and where to go from there. Many who have travelled this path say that the extra effort to get the degree has paid off by taking their career to the next level.

An MBA can help industrial researchers to move to a higher position and earn more. Jane Rhodes, now a manager for new high-tech initiatives at Biogen, a biotechnology company in Cambridge, Massachusetts, had spent ten years at the company working on drugs for neurological disorders such as Parkinson's and Alzheimer's disease. She felt hemmed in by the lab, but she realized that she didn't have the business or management skills to move up the company ladder. "I came through the British education system, which is very focused," she says. "I wanted to learn more about the business side of biotech."

To fill that gap, Rhodes embarked on a two-year MBA programme at Babson College in Wellesley, Massachusetts. Specifically designed for mid-career professionals, the programme took up to 30 hours a week, a big commitment for a researcher who already had a full-time job and a family. The programme would have cost her about US\$75,000, but Biogen paid the bulk of the tuition bill, a sign of how much the company values the degree and the person.

Rhodes used her MBA to get her job at Biogen overseeing new company initiatives, a position that would have been off-limits without the extra training in the business side of science. "I can now move to multiple different positions across the company," she says. "The combination of PhD and MBA is very valuable." She enjoys thinking beyond the confines of research — and that's only one benefit of her revitalized career. "Without an MBA," she says, "I don't know if my salary would be anywhere close to what it is now."

An MBA could give industrial researchers the insight they need to help turn a business around. Looking back, Oréda Boussadia wishes that she'd had that insight in addition to her research skills. She was one of only a few people in the world who knew how to create a certain type of transgenic mouse, thanks to her PhD and postdoctoral training in France and Germany. But she knew nothing about turning mice into profits, which was a problem at the small French biotech company that she joined after her postdoc. "We had very good results, but we had trouble making sales," she says. The company failed within a year, forcing Boussadia to quickly ponder her next step. "I really wanted to continue in biotech, but I had to refine ►

ALUMNI

Post-PhD careers

Most former postdocs from the University of California, San Francisco (UCSF), continue to work in the scientific research enterprise, according to an analysis published earlier this month (E. A. Silva *et al. PLoS Biol.* **14**, e1002458; 2016). The study tracked 1,431 people who left postdoc positions at the university between 2000 and 2013 and had worked in labs supported by the US National Institute of Health's T32 funding scheme. Of the 899 postdoc alumni who did not also have a medical degree and who took jobs in the United States, 81% went on to work in research or teaching, with 336 of those in faculty or faculty-like positions. Another 12% of this cohort work in positions such as policy, communication, regulation, administration and business development.

Around one-quarter of the tracked postdoc alumni went on to work in other nations, and just over half of those gained faculty positions in research or teaching. UCSF postdoc alumni with both an MD and a PhD were also more likely to work in faculty positions than in non-faculty positions, either in or outside the United States.

Employment outcomes also varied by the UCSF labs in which the postdocs worked, although the authors caution that the numbers were too small to be conclusive. Of 49 UCSF faculty members that each served as a mentor for at least 10 postdocs, rates for alumni moving on into faculty positions ranged from a low of 9% to a high of 93%, with a median of 43%.

A paucity of data about where PhD graduates work after their training is often cited as a hindrance to designing more effective employment training programmes. The study authors suggest that institution-based research is necessary to produce data that are sufficiently fine-grained to be useful.

A separate study in *Science* finds that around 40% of US PhD graduates in chemistry, physics or the life sciences think that there is a severe lack of information about non-research careers (H. Sauermann and M. Roach *Science* **352**, 663–664; 2016). The study examined responses from nearly 6,000 US PhD students across 39 institutions, and found that those who said that they had thought at length about their future careers were less likely to decide to do a postdoc. Evidence that postdocs are likely to be default or 'holding pattern' positions points to a need for better career-planning services for graduate students, the authors say. ■

PERFECT TIMING

When to go for an MBA

Timing matters for junior researchers who see an MBA in their future. Although you don't need a PhD to enrol in a programme, many scientists have found that it pays to finish their research training first. "Having a PhD makes it easier to get accepted into an MBA programme," says Jane Rhodes, a director of new initiatives at biotech firm Biogen in Cambridge, Massachusetts. "And non-PhDs who get an MBA have been less successful."

Linh Gilles, director of admissions for the Carlson School of Management at the University of Minnesota in Minneapolis, confirms that applicants to the school's MBA course who already have PhDs are

more likely to be accepted. Recruiting more PhD scientists to the school is a priority, she says. "Students with a research background have that analytical component," she explains. "It allows them to hit the ground running that much more quickly."

Rhodes says that PhD holders who are interested in an MBA should get some industry experience first. "I wouldn't recommend doing it straight out of an academic postdoc," she says. "You have to have some sort of business context." And, as was true for her, scientists who already work in industry might be able to get their employer to pay for some or all of the tuition. **C.W.**

► my management skills," she says. "I knew how to design a research project, not how to develop a company."

Boussadia jump-started her career by enrolling in the MBA programme at the Institut Français de Gestion in Nantes, France. Like other MBA schemes, it focused on the practical aspects of business: product development, market analysis, pricing and return on investment, using real-life examples as learning tools. Degree in hand, she soon got a job managing the production and sales of transgenic mice at a branch of Charles River Laboratories in Lyon, France. After holding that job for five years, she is now the European head of business development and strategy for EpiVax, a biotech company in Lyon. She's happy with the course of her career. "I enjoyed research, but it wasn't enough," she says. "I wanted to be a decision maker."

"I enjoyed research, but it wasn't enough. I wanted to be a decision maker."

NEW HORIZONS

Armed with an MBA, many can leave the lab without leaving science. As a postdoc, Kyle Rasbach investigated potential therapies for muscular dystrophy at the Dana-Farber Cancer Institute in Boston, Massachusetts. But thanks to the MBA that he'd pursued along with his PhD, he was snapped up after his postdoc for a job studying investment opportunities at investment management firm T. Rowe Price in Baltimore, Maryland. Much of his remit involves evaluating the research taking place at drug companies, from the giants of the business to small start-ups. His lab background helps him to spot blockbuster drugs in the making. "Sixty to seventy per

cent of my job is science-based," he says. "You can't do this job and be excellent at it without a PhD or an MD."

That's also true for Moritz Fischer, director of international marketing for Fresenius Medical Care in Hessen, Germany. After earning his medical degree at the Ludwig Maximilian University of Munich in Germany, he realized that he did not want a career as a physician or clinician. He took a job at Fresenius as a lower-level marketing manager, but soon recognized that he could go much further with advanced business skills. So he pursued an MBA at Danube University Krems in Austria. The company covered his tuition, which he estimates would have cost him at least €20,000 (\$22,500). It was a reasonable investment for the company, he says, because he has made money for them. "They were able to capitalize on my training," he says.

Success stories of researchers with MBAs in biotech and drug development have caught the attention of early-career researchers who are still plotting their careers. Jeffrey Zahratka, a postdoc at the Cleveland Clinic in Ohio, says that he could see himself working at a biotech firm, perhaps one that makes implantable devices to treat neurological disorders. "I could act as a go-between for the research side and the business side," he says. He still has to weigh up the pros and cons of another degree, but he thinks that he could bring a lot of value to a company. "People with a research background have a lot of tenacity," he says. "They are battle-tested."

If he decides to go down the MBA route, he won't be alone. But for now, PhD-MBA remains a relatively rare combination — that factor alone can help a person to stand out and move forward. It's a matter of degree. ■

Chris Woolston is a freelance writer in Billings, Montana.

BROWSING

Security issues.

BY IAN WHATES

A slow day. Gabriel 'Gabe' Tarvy found himself on the corner of 52nd and 3rd. It was nowhere special — exactly the sort of place he favoured when browsing.

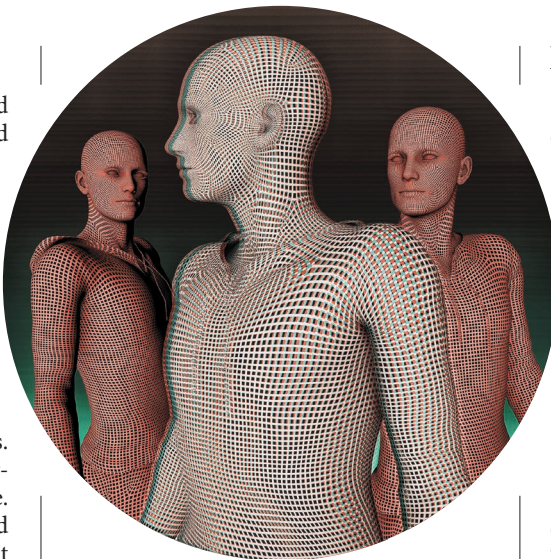
The intersection was busy — not jammed but with more than enough foot traffic to provide cover. Gabe loitered at the corner, allowing his gaze to sweep this way and that as if to get his bearings. In a handful of seconds, twenty-eight people had passed close enough to be registered and identified by his smart lenses. Twenty-five were of no interest — their personal security was too tight, too up-to-date. Gabe felt confident he could have cracked any one of them given time, but not without triggering alarms, not unless he had a *lot* of time...

The other three, though, they were a different matter.

This was how it went, how it always went. He just needed to be patient. Wait long enough and someone who hadn't yet applied the latest security updates was bound to come along. They intended to upload them, no doubt — tonight, or maybe tomorrow — but they hadn't got around to it yet. Sometimes, as today, he didn't have to wait long at all.

Two of the three were exactly what Gabe had been hoping for: low-grade security, outmoded and ineffectual if you knew what you were doing. The first he dismissed. The clothes were shabby, the security so cheap and inadequate that, rather than being late to upgrade, it was clearly all the man could afford. Not worth bothering with. The other, though, was well-dressed and professional-looking; lazy rather than being poor, just begging to be fleeced. Gabe took great delight in doing so, stealing passwords, plundering accounts, all in the two or three seconds it took the mark to stroll past. The funds disappeared via a series of transfers and switches between dummy accounts registered all around the world, before eventually returning to Gabe, scrubbed and untraceable.

The third anomaly was intriguing. A woman, early thirties, professional and sharp, sporting the sort of suit that said she was going places if she hadn't already arrived. Everything about her screamed money, particularly her security, which was better than anything Gabe had encountered



while browsing before. This sort of sophistication he expected to find guarding the core secrets of a major corporation, not an individual strolling along 52nd Street.

He reined everything back, wary in case his system's lightest touch should trigger an alert, and then he followed her. Of course he did. Gabe loved nothing more than a challenge. Once he established where he might find her again he would return with subtler, more sophisticated tools. Then he would have her.

The crowds made tailing her easy. She led him to a smart office block, the sort occupied by numerous companies or by one corporate giant. She entered via large plate-glass doors. After the slightest hesitation, Gabe followed. If he could just see which floor she went to, work out who she worked for... But a security desk stood between him and the elevators. Anyone passing beyond would be noted, scanned and challenged if they didn't belong.

This would be enough. He could come back better prepared and wait for her outside.

As he turned to leave, Gabe was confronted by two burly men. Before he could reach the woman was there, penning him in.

"Gabriel Tarvy, come with me, please." The words were clipped, her voice assured.

"What do you...?" He started to protest, but she skewered him with a look that caused the words to wither on his lips.

"Do you really want to do this in public?"

He had no answer. "I didn't think so. This way."

He followed meekly to a side office, the

two goons in suits never more than a step behind. At least they stayed outside.

"What is all this?" Gabe asked, recovering some of his customary confidence.

The office was sparse, bright, antiseptic. There was no desk, just two chairs facing each other. The woman gestured towards the nearest. Gabe sat, reckoning the sooner he did so, the sooner this would be over.

"My name is Laura Dyne," she said, "and I'm here to recruit you."

This was ridiculous. "To do what, exactly?"

"To catch other Browsers."

"You're police, then?"

Her smile was thin. "No, we're private sector, contracted to keep the streets safe from casual thieves like you. Have you any idea how many billions browsing costs the economy every year?"

Gabe did, but saw no reason to admit as much. "Why should I?" he said. "You've got nothing on me."

"The funds you just stole from the man with the low-grade security, they were marked. We can follow their every movement, their every transfer, which will lead us back to..."

Him. But that was impossible. Transfers couldn't be traced... Could they?

"It's up to you, of course," she continued.

"So there is a choice."

"There's always a choice." He didn't like *that* smile, it was unnerving, predatory. "Either you join us or we Black Flag and incarcerate you."

"No!" The denial was out before he knew it. Prison he could handle — the sentence wouldn't be long, not for a first offence — but Black Flagging... It meant being indelibly tagged: nanotech, binding with his very DNA, a stain that could never be excised. He would be marked. Anybody with even the crudest personal security — which meant everyone — would recognize him as dangerous, to be avoided.

It meant becoming a pariah: no work, no friends, no anything.

"You wouldn't," he said. But he saw in her eyes that she would.

What she offered him was no choice at all. ■

ON NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

Ian Whates is the author of the Noise books (*Solaris*) and the City of 100 Rows trilogy (*Angry Robot*). His second collection, *Growing Pains*, came out in 2013.

ILLUSTRATION BY JACEY